DOCUMENT RESUME

ED 151 316 95 SP 012 276

TITLE	Teaching-Learning Interaction Study. Progress Report Through September 30, 1977. Condensed Version.
INSTITUTION	Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	Sep 77
CONTRACT	NE-C-00-3-0066
NOTE	148p.; Page 91 will not reproduce clearly due to small type
EDRS PRICE	MF-$0.83 HC-$7.35 Plus Postage.
DESCRIPTORS	*Academic Achievement; Interaction; *Interaction Process Analysis; *Learning Processes; Performance Factors; Student Attitudes; Student Characteristics; *Student Teacher Relationship; Teacher Behavior; Teacher Characteristics; Teacher Education; *Teaching Methods

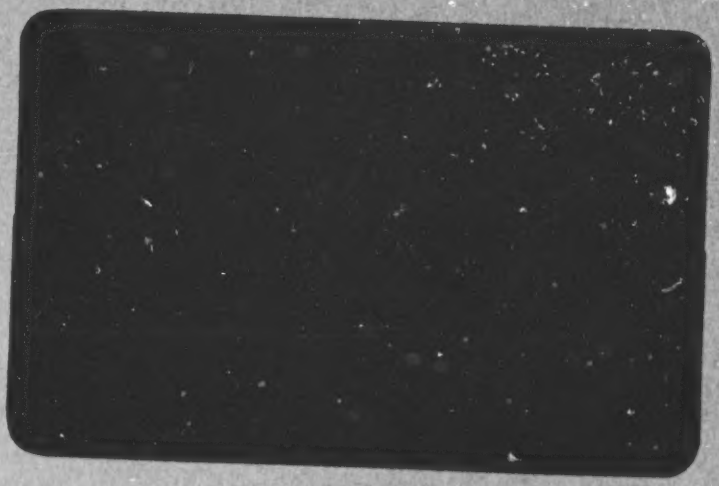
ABSTRACT

This is an interim progress report on an ongoing study of teaching-learning interaction. The function of the study is to observe learning over a full year of schooling, focusing, in part, on reciprocal interactions. The teaching-learning interaction program is committed to three long-term objectives. Methodologically, it is developing improved instruments and new statistical models that make it possible to assess the interacting effects of teacher-student characteristics on multiple student outcomes: academic progress, attitudes toward learning, coping skills, and self-esteem. Substantively, it is looking at the specific effects of different teaching strategies on different kinds of students, including differences due to socioeconomic level and ethnicity, as well as individual characteristics. Practically, it aims to find out whether it is possible to develop appropriate prescriptions for achieving optimum effect on a specified outcome for a particular kind of student. A tentative conclusion is reached at this particular point in the continuing research on this subject. It is suggested that true relatedness does exist between teacher personality and teacher behavior; between child personality and aptitude and child behavior; and between the interplay of teacher and child, and the changes in student knowledge, coping skills, and attitudes that take place over the school year. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED151316

Teaching-Learning Interaction



The Research & Development Center for Teacher Education

The University of Texas at Austin

UTR&D

SP012 276

TEACHING-LEARNING INTERACTION STUDY

Contract NE-C-00-3-0066

PROGRESS REPORT

through September 30, 1977

Robert F. Peck, Director

CONDENSED VERSION

The Research and Development Center for Teacher Education
The University of Texas at Austin

September, 1977

This is an interim progress report on an ongoing study supported in part by National Institute of Education Contract NE-C-00-3-0066. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that office should be inferred.

Please do not quote or cite this report without the written permission of the project director.

AN INTRODUCTION TO THE CONDENSED VERSION

This is a distillation of an 861-page report. Even so, many readers will prefer to skim it, just hitting certain highlights that meet their interests. It may be helpful to sketch a "table of contents":

Overview 1-7

Design 9-20

Propositions to be tested . . 21-36

Instrumentation. 37-59

The Sample 67-70

Analytic Plan 71-86

Some preliminary findings . . 87-197
(many pages omitted)

Future Research Extension . . 218-219

All page numbers shown are those of the original full length document. For this version, however, only a few selected pages are printed here. The rest are omitted.

TEACHING-LEARNING INTERACTION STUDY

TLIS PROGRESS REPORT

INTRODUCTION TO TLIS

OVERVIEW

Experienced teachers are keenly aware that it often takes different approaches to achieve equivalent learning results with different students. There is no one "right way" to teach everyone. Effective teaching is as much a function of the needs, readiness and reactions of the widely varying students as it is of the teacher's strategy. In short, most learning is the product of an interaction between student and teacher characteristics; neither alone can accurately explain learning.

Teachers also know that it is not enough to look only at a test score, or at some other measure of academic learning, to evaluate the important outcomes of schooling. It is equally necessary to look at the effects on students' attitude toward further learning, on their self-confidence and self-respect, and on their generalized work-habits and coping skills. Multiple outcomes of education must be assessed simultaneously before one can judge whether a year's schooling has been good or bad for the students, and in precisely what ways.

These propositions seem so self-evident that children, parents and teachers almost automatically nod their heads when they hear them stated. They see at once that research should follow this logic, so that accurate prescriptions may be discovered for fostering the intellectual and emotional growth of the many different kinds of students who come together in school. They realize that a teacher's individual nature may make some approaches easier or more congenial for him or her. They recognize that no teacher can reasonably be expected to master every style of teaching, nor to get outstanding results, in every way, with every kind of student.

In recent years, major systems of instruction have been developed to try to take this into account: Individually Guided Education, Individually

Prescribed Instruction, Personalized Teacher Education, Keller's Self-Paced Method, and others.

Even these systems involve much more art than science. They rely almost entirely on intuitive teacher judgments, or on untested assumptions about what works best for whom, for there is very little research-based evidence, as yet, about what is most effective for different kinds of students. Indeed, as we grapped with this issue at The University of Texas R&D Center over the years, we found that the statistical technology had not yet been developed to identify accurately the different effects of instruction on different students when even as few as two kinds of input variables need to be considered simultaneously.

As Cronbach (1975) pointed out, a simplistic search for "general laws" of human behavior is almost certainly either useless or incorrect at this stage of our knowledge. What is needed are studies that would allow us to say, for example, that instructional method X will produce intellectual progress for student type A, while maintaining his/her interest in further learning; whereas method Y will maintain interest but not promote much learning; while method Z will promote learning but kill interest. At the same time, student type B may learn enthusiastically from method Y, not learn as much from method X, and will have both learning and interest killed by method Z. In short, what is needed is research that reasonably reflects the multivariate nature of real-life learning and the interacting effects of teacher and learner characteristics as they affect important outcomes.

To study these problems, and to design more effective schooling, requires a logical model that examines the interacting effects of teaching strategies and student characteristics on multiple outcomes of instruction. In a growing number of studies, the use of an interaction paradigm has demonstrated that learning outcomes can only be explained as the result of

an interplay of forces. Studies that fail to allow such interactions to be measured simply cannot arrive at an accurate map of the complex world of everyday learning. There are problems aplenty in this approach. Even in the course of updating his argument for a multivariate approach to the study of behavior Cronbach (1975 , 1976) points to troublesome limitations in our best, present quantitative methods for handling such issues. Nonetheless, he sees almost no validity at all in more simplistic approaches that ignore interaction effects, curvilinearity, or other evidence which indicates that not all people react the same way. He has found serious conceptual and statistical difficulties which must be overcome in any analysis which uses data on individual children, pooling classes. (Cronbach, 1976). (Corrections for these factors have been made in the present study by selecting a nearly randomized sample, and by testing for "class" effects in assessing the effects of teacher characteristics. See Analysis Section below).

Hunt (1975) presents a recent review of the evidence and the logic for what he calls the Person-Environment-Interaction (P.E.I.) paradigm. He proposes four essential elements for such a model: (a) that it look at multiple behavioral outcomes; (b) that it be "developmental" over some appreciable period of time; (c) that it look at the reciprocal interactions between person and environment; and (d) that it be "practical."

The present study does look at multiple outcomes; it observes learning over a full year of schooling; it specifically focuses, in part, on reciprocal interactions; and it attempts to be practical, both by looking at learning in its natural setting, in school, and by assessing outcomes that are widely held to be important. This line of research at The University

of Texas grew out of the multidisciplinary tradition of the Committee on Human Development of the University of Chicago, represented by the work of such people as Havighurst, Warner, and Prescott. That tradition embodies the same debt to Kurt Lewin that Hunt acknowledges. It has shaped the design and instrumentation of a series of research studies, starting with the Mental Health in Teacher Education project (Peck, 1958); proceeding through the Teacher Personality, Teacher Education and Teaching Behavior study (Peck, 1962); the Computer Analysis of Personality project (Peck and Veldman, 1962); the program design for the R&D Center for Teacher Education (Peck, 1965); its Personalized Teacher Education program(Peck, 1970); its Individualized Teaching for Effective Coping project (Peck, 1968); and the Cross-National Study of Coping Styles and Achievement (Peck, 1972).

A considerable number of studies have used some form of the interaction paradigm. As early as 1944, Thompson and Hunnicutt reported that "introverted" students performed better when praised by their teachers, whereas "extroverts" performed better when criticized by teachers for their mistakes. Heil (Washburne and Heil, 1960) reported that teacher personality interacted with child personality to create differential learning in different types of children, with different patterns for different elementary school subjects.

Cronbach (1957) strongly recommended that an interaction paradigm be adopted as the best logical model for tracing educational effects to their actual causes because this better approximates the complex realities of everyday learning. Subsequently, quite a few studies attempted to embody this approach, although many of them used a very narrow definition of "aptitude-treatment interaction" (ATI). Moreover, according to Bracht (1970), "the analysis of an interaction effect was often an afterthought rather than a

carefully planned part of an experiment." Nonetheless, Glass (in Wittrock and Wiley, 1970) concluded, "ATI has not paid off."

On the other hand, Cronbach and Snow (1969)found that the interaction approach explained learning better than the single-predictor model. Lesser (1971) cited a substantial number of other studies in support of this approach. Berliner and Cahen (1973) did not find as powerful confirmation of the trait-treatment interaction model as they expected, but they did see hope for it in better-conceived studies. Koran (1971), reporting a study with numerous interaction effects, observed "as Cronbach and Gleser have suggested (1965), initial study of aptitude X treatment interactions will quite possibly be more important for what it tells us about the psychology of instruction than for immediate placement purposes. However, such experimentation may eventually help provide a basis for the individualization of instruction."

Britt (1971) described an intriguing, computerized method of identifying "learner types" which, from the defining statements, also specifies the characteristics of differentiated instructional programs that should optimize the learning of each "type." He did not, however, cite evidence of the observed effectiveness of such differential instruction in practice. Salomon (1971) reviewed a number of studies that showed important ATI effects. His 1972 paper describes ways in which ATI can be turned into specifications for differential programs of instruction for different kinds of learners (Salomon 1972).

Studies reporting significant trait-treatment interaction effects that have not been cited in earlier reviews include the following: Berliner, 1972; Blitz and Smith, 1973; Britt, 1971; Brophy, 1975; Cronbach and Snow, 1969; Davis, 1967; Dowaliby and Schumer, 1973; Featherstone, 1973; Fisher,

1973; Kress and Gropper, 1966; Lippman 1970; Pervin 1968; Shores, 1969; Smith, Wood, Downer, and Raygor, 1956; Tallmadge and Shearer, 1971; and Taylor, 1970.

It is easy to sympathize with the critics of the P.E.I. approach. In struggling to use it in various action-research and pure research studies here at The University of Texas over the past twenty years, we have found that it is difficult and expensive to conceptualize this approach sharply, to instrument it soundly, to carry it out in practical settings, and to analyze it. Some of the necessary statistical procedures, for example, have only been developed into a usable form within the past four or five years. Some of the key issues are not yet settled. Nonetheless, every time we have used this approach it has proved out. The preliminary findings from the present study seem to confirm that it is every bit as essential to an accurate, insightful, useful analysis of human learning and human development as we initially supposed it might be.

The Teaching-Learning Interaction program is committed to three long-term objectives. Methodologically, it is developing improved instruments and new statistical models that make it possible to assess the interacting effects of teacher-student characteristics on multiple student outcomes: academic progress, attitudes toward learning, coping skills, and self-esteem. Substantively, it is looking at the specific effects of different teaching strategies on different kinds of students, including differences due to socioeconomic level and ethnicity, as well as individual characteristics. Practically, it aims to find out whether it is possible to develop appropriate "prescriptions" for achieving optimum effect on a specified outcome, for a particular kind of student. Any such "prescriptions" would thereafter need to be field-tested, so that those that proved valid could be incorporated

into training materials for in-service and pre-service teacher education.

The design provides for double testing of all findings by replicating the study with the same teachers but different classes in two successive years, and by replicating this plan in Austin, Texas, and Daviess County, Kentucky. The instrumentation provides reliable measurement of teachers' personal characteristics, their teaching behavior, and the multiple student characteristics, demographic as well as personal. The Austin sample consists of a representative sample of the Anglo-Black-Chicano population, randomly assigned to ungrouped classes in integrated, sixth grade learning centers. A 9-stage analytic model has been developed which tests for linear, curvilinear, and interaction effects of teacher and student variables, and a test for the appropriate unit of analysis (individual student or class) to correct for possible nested variance. "Gains" are assessed by covariance, using the pretest as a covariate, since this is the soundest way of estimating changes and its correlates. Findings that replicate across the two years and the two regions should have quite strong generalizibility to similar populations.

Along with the quantitative data, considerable descriptive data are being collected during the observations and in repeated interviews. This yields brief but behaviorally detailed case studies of the interaction process between each teacher and a varied sample of individual students.

Ways are being developed to give useful, understandable feedback to the participating teachers and administrators. This includes overall findings and findings on an individual class that can be given privately to the teacher.

CONCEPTUAL DESIGN AND INSTRUMENTATION

Figure 1 presents a simplified conceptual system, adapted in part from Bloom (1956) and Krathwohl (1964). It identifies three kinds of skills: cognitive, affective, and coping; and three aspects of life to which each skill applies: career-related activities, interpersonal activities, and intrapersonal experiences. The specific skills listed within each of the nine cells are only a few illustrations, out of many that could be described; but they are chosen for their saliency for education and for living.

When looked at as goals for students , they specify desirable outcomes that parents and educators have long assumed should be the goals of public schooling. When looked at with teachers in mind, they describe skills which are desirable for teachers to possess, in order that they may facilitate the growth of their students in these same directions.

Figure 2 identifies the instrument-variables that were selected or designed for this study, in order to provide operational measures of some of the skills/goals in each of the cells that were explained in Figure 1. Thus, a pupil's knowledge of facts, and some measure of analytic skill, are assessed by a standardized achievement test. A pupil's or teacher's accuracy in identifying and using information is assessed by the observational code (COS-Accuracy). A teacher's feelings and empathy for students are assessed by self-report (ASD-Behavior), by observer rating (COR-KU), and by observer codes of behavior (COS-Affect); etc.

The pupil and teacher assessment data are further organized into a network of cross-linkages. Figures 3 and 4 describe the presage measures, taken four to six weeks into the school year; a set of process measures based on periodic observations and interviews through the year (low, medium or high inference measures, in different instances); and outcome measures, taken in April, the

GOALS OF EDUCATION

(Applicable to Teachers as Well as Students)

	Career Related Domain	Interpersonal Domain	Intrapersonal Domain
Cognitive Skills	Knowledge of facts Knowledge of ways of dealing with specific intellectual tasks Knowledge of systems for conceptualizing, ordering, manipulating, processing, studying Ability to analyze problems Ability to create new solutions	Knowledge of others' individual natures and dynamics Knowledge of social skills and customs Understanding; ability to analyze social situations, dynamics; awareness of how others see oneself Ability to perceive others' thoughts, emotions needs Ability to differentiate social situations, personal and cultural styles and determine appropriate interaction style	Accurate self-awareness of mind, body, feelings, actions Awareness of own values, moral principles, goals, and their coherence Awareness of strengths, weaknesses of character, personality Knowledge of ways of coming to self-realization, of expressing oneself and evaluating self Mapping a clear career-path (future plans, awareness of own strengths, weaknesses, and needs)
Affective Skills	Taking pride in doing work well; a sense of responsibility for creating a good end result Valuing learning and knowledge, in general; having an active interest in some specific fields of knowledge Valuing the "truth ethic", honesty, wanting to find truth Feeling comfortable in work-focused settings Developing an emotional commitment to a career	Sense of others' feelings: empathy Positive attitudes toward others Having a flexible, versatile repertoire of realistic feelings toward others Valuing human intimacy Caring about others' well-being; altruism; a sense of moral responsibility Acceptant of constructive superior-subordinate and co-worker relationships	Self-esteem: attitudes about self, including one's performance in all cells: evaluation of accuracy of perceptions; correctness of values; comfort or discomfort with one's feelings; adequacy of knowledge or social skills; etc. Integration of "truth ethic" into self-evaluation and into one's choice of actions Achieving reasonably coherent, harmonious feelings about the major issues in one's life Developing a coherent, personal system of values and goals Developing spirituality; a religious or cosmic sense of self/Other relatedness
Coping Skills	Developing good work habits: vigorous, unstinting, goal-directed effort Pursuing a coherent career-plan (whether consciously or intuitively) Increasing precision, efficiency and integration of career skills, through self-analysis, self-correction, and purposeful improvement Increasing one's insight into the "larger picture" into which one's own work fits: the many different problems involved and skills required to make the whole system work; putting these insights to work Developing the intellectual and emotional flexibility to change careers if it becomes either desirable or necessary	Responding with interest and empathy to others' feelings, interests and values Responding with open-minded attention to others' ideas and suggestions Reconciling one's own and others' wishes, to achieve a mutually tolerable outcome Fostering others' self-esteem Dealing with others in a trustworthy, caring, ethically responsible way	Ability to acknowledge and express emotions, to respond to events and people sincerely Ability to act on one's values and principles Ability to care for one's own real physical, emotional, and intellectual needs Ability to cope with anxiety Achieving a well-balanced personality: congruent in thought, feeling and action

Figure 1

second-last month of the school year. As much independence of measurement as possible is built into the system by using several sets of informants about the same issue (teachers, students and observers). This permits multiple cross-validation of any one measure, in up to seven ways (see Psychometric Analysis: Validity, below).

Since the measures are taken in a three-stage temporal order, they can be analyzed for antecedent-subsequent relationships in ways that imply causal explanations. Even though it cannot prove causality, such analysis can greatly sharpen and focus attention on particular relationships which can thereafter be put to experimental test for rigorous proof of causal connections between antecedent and subsequent phenomenon.

Many, if not most of the choices and arrangements of instruments are designed to test a series of general and specific hypotheses, as described just below. In addition, such a complex network may also lead to quite unanticipated findings. For example, Ryans' COR method was chosen for two reasons. First, it starts with medium-inference rating scales that have considerable specificity, and therefore should be reliable (they are, in this study, as in his original research). From these scales, however, are derived high-inference factors that seemed to us very likely to generalize across occasions -- to represent stable dimensions of individual teacher behavior. Second, Ryans' Systematic-organized factor seemed likely to have a direct effect on student Time-On-Task, and on student achievement and coping skill outcomes. His Kindly-understanding factor seemed likely to have a positive effect on student self-esteem, affect in class, and attitude toward school. Thus, for both psychometric and substantive reasons, this looked like a good observation system to use, as opposed either to one with highly specific, low inference codes or one with extremely high inference judgments such as Kounin's "withitness," vividly though

TLI INSTRUMENTS CLASSIFIED BY SKILL DOMAINS

	Career-Related		Interpersonal		Intrapersonal	
	Pupil	Teacher	Pupil	Teacher	Pupil	Teacher
COGNITIVE	Achievement Tests COS-Fluency, Accuracy Unit Tests	COR-SI COS-Appropriate Level COS-Accuracy		Interviewer Perceptions: Diagnostic, Prescriptive Skills	BRS-SRx BRS-OR BRS-SRx Stud.Descrip. and similar comparisons to assess accuracy of self-perception	ASD, VOTxCOR, etc. to assess accuracy of perception of different aspects of self
	CORF anecdotes Teacher Interviews	CORF Teacher Interviews	CORF Teacher Interviews	CORF Teacher Interviews	CORF Teacher Interviews	CORF Teacher Interviews
AFFECTIVE	SSI SSC-A SET (indiv.score) COS-Affect	VOT-Attitude	SSI SSC-A	ASD-Behavior, Individualism, Introversion Concerns Level COR-KU COS-Affect COR-P (class score)	P-H, SSD How This Class Makes Me Feel (individual score)	ASD-Attitude ASD-Attractiveness
	CORF Teacher Interviews	CORF Teacher Interviews	CORF Teacher Interviews	CORF Teacher Interviews	CORF Teacher Interviews	CORF Teacher Interviews
COPING SKILLS	BRS-OR SSC-C Student Description COS-TOT, Act/Pass. Class Means COR-P Class Checklist CORF Teacher Interviews	VOT-Coping ASD-Efficiency COR-SO COS-TOT Interview: Follow-Through SET(class mean) CORF Teacher Interviews	BRS-OR,SR. Student Description CORF anecdotes Class Means COR-P Class Checklist CORF Teacher Interviews	How This Class Makes Me Feel (class mean) CORF Teacher Interviews	BRS-OR,SR Student Description CORF Teacher Interviews	ASD-Anxiety CORF Teacher Interviews

Figure 2

that conveys a potent concept. This design therefore makes up somewhat in exploratory power for the power it lacks to furnish final, definitive proof of the kind that a univariate or bivariate experiment can supply -- when it is founded on sound intuitions from previous, more exploratory investigations. The present study is thus a hybrid of hypothesis-testing and hypothesis-generating efforts.

There were a number of crucial problems to be faced, in designing the study. The problem of nested variance, where students are not at all randomly assigned to schools or to classes, in many communities (Mayeske, 1970), could have been a fatal obstacle to using the individual student as the unit of analysis, which the interaction model requires. This problem was largely resolved by careful selection of the sample. Austin presently operates sixth-grade learning centers where the students are carefully mixed to represent the city's ethnic ratio, and randomly assigned across classes. Our later tests for "class" effects -- see Appendices B-6 and B-8, indicated that there was remarkably little non-random sorting of students compared with the situation in many other schools. The same was true of the teachers, who were not "matched" with their classes to any strong degree; on most variables, not at all. The analyses of the data from this study verify that the school administration's policy was, in fact, carried out to a very great degree. Finding this, makes it possible to go ahead with tests of the main propositions of the study.

In Daviess County, randomization of the almost 100% Anglo population was not a political issue, and the heterogeneity of the school population by socio-economic indicators seemed high, to informed observers. The tests for nested variance have yet to be made on those samples, so it is too soon to know how they will turn out.

TLI INSTRUMENTATION DESIGN

	Presage Characteristic Measures (Pretests)	Process Measures (All teachers and Special Study Students, except where noted)		Outcome Measures (Posttests)
		Behavior Measures	Behavior Ratings	
TEACHERS	BIO: Ethnicity, Sex, Age, SES of Family of origin, Years of Teaching Experience; Degree Attained DI: (some cases) ASD: Attitude, Behavior, Efficiency, Introversion, Anxiety, Individualism, Attractiveness VOT: Coping and Attitude SET PRE: class mean Concerns Level	Year 1 CORF: descriptive data COS: source of CORF material Year 2 COS: TOT; Affect; Accuracy; Appropriate Level; Source of CORF material CORF: descriptions Interviews: descriptions	Year 1 COR - KU, SO, SI Year 2 COR - KU, SO, SI Interviewer Perceptions: Diagnostic, Prescriptive, Follow Through Skills Scales	SET: class mean HCMF: class mean (year 2 only)
STUDENTS	BIO: School, grade, class, ethnicity, sex, age, SES Achievement: standardized test (CTBS, MAT, GMG) Coping Skill: SSC-C, BRS-OR,SR Attitudes: SSC-A, SSI, SET Self-esteem: P-H (SSD)	Year 1 COS: TOT CORF: anecdotal material Final Interview: teacher remarks on Special Study Students Year 2 COS: TOT; Accuracy: Rd.Fluency; Affect; Active/Passive CORF: Anecdotal material Teacher Interviews: Teacher remarks about students Achievement: Unit Tests (Austin only, all students)	Student Description: teacher rating of Special Study Students on BRS and COR P items) SD - Status: Year 1 and Year 2 SD - Change: Year 2 only; how has student changed over the school year? Attendance	Achievement: standardized tests Coping Skill: SSC-C; BRS-OR,SR(Year 2 only) Attitude: SSC-A, SSI, SET Self-esteem: P-H (SSD); HCMF(Year 2 only)
CLASS	Achievement: class means on standardized tests Coping Skill: class means on SSC-C, BRS-OR,SR Attitude: class means on SSC-A, SSI, SET Self-esteem: class mean on P-H (SSD)	CORF: class descriptions COS: class average TOT, Accuracy, Rd.Fluency, Affect, Active/Passive Achievement: reading group means on Unit Tests	(Full Class) COR-Pupil factor Class Checklist - Similarity Index	Achievement: class means on standardized Tests Coping Skill: class means on SSC-C; BRS-OR,SR (Year 2 only) Attitudes: class mean on SSC-A, SSI, SET Self-esteem: class mean on P-H (SSD)

Figure 3

Even in Austin, the selection of the sixth grade centers was not without its price. In these schools, each class of students rotates during the day among three or four teachers. We picked the teacher who had the class for homeroom, as well as for the social studies or reading period (in most cases), so that this teacher presumably had twice the chance to influence the students as any other one teacher. This is not a critical problem where the student outcome measure was tightly linked to that teacher, such as the SET, the achievement test or, in Year 2, the How This Class Makes Me Feel measure of that teacher's impact on the student's self-esteem. The more general outcome measures of attitude (SSC-A, SSI), self-esteem (P-H), and generalized coping skill (SSC-C; BRS-OR POST, in Year 2 only) were not that tightly linked, however. It is almost a wonder that any systematic "teacher effect " is found, on these measures, considering all the extraneous "noise" in the system: the other two or three teachers, especially.

Indeed, even the achievement test items are so extremely general in their content, and so distantly related (if at all) to the content actually covered in "the teacher's" class, that it seems to us they are among the least valid of all the measures for attributing a student outcome to a particular teacher. They appear to be an honest assessment of how much a student knows about a given kind of test item, at any time, pre or post. What that knowledge is due to, though, can and probably does include everything the student learns from informal conversations and adult examples at home, from television, from older brothers and sisters, from his recreational reading (or lack of it) -- everything that either fosters or inhibits learning.

Indeed, one of the greatest unmet needs, at the outset of the study, was for sturdily reliable, valid instruments to assess the input and outcome variables. Through a combination of search and invention (see Appendices

TITLES, ABBREVIATIONS AND ADMINISTRATION PERIODS OF INSTRUMENTS

	Title	Abbreviations	Year 1 fall	Year 1 spring	Year 2 fall	Year 2 spring	Notes
TEACHERS	Teacher Biographical Form		x		x (new teachers only)		
	Adjective Self-Description	ASD	x		x (new teachers only)		
	Views of Teaching, Attitude and Coping Scales	VOT -A,C	x		x (new teachers only)		
	Directed Imagination	DI	x				(Not all teachers completed DI)
	Student Description	SD		x		x	(Ratings on Special Study students)
	Class Checklist			x		x	
	School Preference Questionnaire			x			
	Teacher Concerns Questionnaire					x	(Austin reading teachers only)
	Teacher Needs Survey			x			(Austin only)
	Teacher Reaction Form					x	
	Class Roster		x	x	x	x	
	Student Attendance Form			x		x	
STUDENTS	Student Biographical Form		x		x		
	Student Sentence Completion, Attitude and Coping Scales	SSC-A,C	x	x	x	x	
	Student Evaluation of Teaching	SET	x	x	x	x	
	School Sentiment Index	SSI	x	x	x	x	
	Behavior Rating Scale -Self-Report Other-Report	BRS-SR,OR	x		x	x	
	Piers-Harris Children's Self-Concept Scale	P-H	x	x			(P-H was modified and renamed)
	Student Self-Description	SSD			x	x	
	How This Class Makes Me Feel	HCMF				x	
	Comprehensive Tests of Basic Skills	CTBS	x	x	x	x	(Yr.1-all students; Yr.2 - some classes)
	Gates-MacGinitie Reading Test	GMC			x	x	(Austin only)
	Metropolitan Achievement Tests	MAT			x	x	(Kentucky only)
	Unit Tests				continuous measures		(Yr.-2 Austin only)
OBSERVERS	Classroom Observation Record: Kindly-understanding, Systematic -organized, Stimulating- inventive factors	COR-KU,SO,SI		x		x	
	Classroom Observation System	COS		x		x	(revised; Yr.2. Yields several teacher/pupil vars)
	Classroom Observation Report Form	CORF		x		x	
	Interviews (initial, weekly, final)			x		x	
	Interviewer's Perceptions Diagnostic Prescriptive, Follow Through Skills	IP-D,P,FT				x	(Rating of teacher's diagnostic, prescriptive and follow through skills)

Figure 4

A and D) a usable set of instruments was assembled. They are being validated more thoroughly in this study than had previously been done, in most cases. Self-report, machine-scorable instruments are scarcely the most valid or powerful measures to use. They meet an extremely practical criterion, however: they are affordable, for use with large populations of students. This was not a crucial consideration for the study, as basic research, but it is crucial if there is ever to be any prospect of using such instruments in everyday public education using scarce public funding. No matter how good a measure might be, if it costs too much, few children will ever get the benefit of it.

This is not to say, of course, that even positive findings with the present instruments would necessarily imply that they should be used systematically in schools to diagnose the qualities they assess. The important findings concern the properties of teachers and students that are found to be related. Who should assess them, in everyday practice, and how, is quite a different and open question. The complete findings from this study may have something to say to that question; it is too soon to address it now.

The most urgent, single condition for this study was to include a large enough number of teachers (classes) to permit this many measures to be used. Although the number of students grew by 25 or 30 with every teacher added (and so did the cost of assessing the children), the minimum sample size was set by the number of teachers. That, in turn, was sharply and firmly limited by the available funding, which was lower than the projected costs, at the outset, and shrank even further over the years as a function of reduced or "level" funding of the sponsoring agency, year after year, while inflation galloped ahead. One practical consequence was that almost all the budget had to go for instrument development and data collection until all the data were in, in 1976. Only one staff member was available to do as much data processing, compilation

and programming as one person could do, in the first two years of the study. This guarantees some errors will slip through that cost considerable staff time, later on. Nonetheless, such decisions did preserve the sample size and kept the data coming in.

Or, rather, it was the heroic efforts of the cooperating teachers and the research staff, who kept working to plan no matter what vicissitudes befell the real-dollar budget for the project. Carrying out a research study, particularly one of this relative magnitude, is no mere exercise in intellectual self-discipline by a solitary thinker. It is a very human story, depending for its success on the patient, dedicated perseverance of a great many people. Such qualities of unselfishness, intellectual integrity and conscientiousness are in rich supply among the people in our schools and universities; more than one might think from the rather acid tenor of the critics of education in the past decade.

From the earliest phases of the study -- indeed, in previous studies beginning in 1968, developing an appropriate, rigorous system for analyzing change in this multiple impact, multiple output design was a vital concern. While cluster analysis sounded logically plausible, we knew both from theoretical grounds and from trying it with earlier data that it did not contain some of the rigorous, precise tests of significance which seemed to us essential. Canonical correlation had some analogous limitations, and also appeared difficult to translate into the specification of particular teaching-learning effects, within an overall pattern. For over a year, Dunivant explored the possibility of econometric models for multiple linear regression; but, while powerful in their realm, they, too, proved to reduce the data far too much, in a way that made it impossible to specify particular antecedent - subsequent relationships in ways that might be useful to educators.

Cronbach early counseled us that a "true score" correction for instrument reliability should be an indispensable part of any statistical system we adopted. Porter and Cohen strongly seconded this idea. That was the state of the art up to 1976. We tried. Indeed, all of the Year 1 data were subjected to such a correction, at one point. Some doubts arose, however, and Marston, then in charge of data analyses for the project, ran a Monte Carlo study with synthetic data. He found that the "correction" actually inflated the number of "significant" findings in a spurious way (Marston, 1977). Cronbach re-examined the issue and also concluded that the available "true score" corrections could increase rather than decrease error (personal communication, 1977). Cohen, principal statistical consultant to the project in 1976-77, simultaneously tested this issue out and reached the same conclusion. Just recently, Dunivant, now at N.Y.U, discovered and worked through Jöreskog's new method of correcting for reliability, and recommended that we try it. This will be done in 1978.

We needed to test for class effects, as Cronbach has persuasively demontrated. We wanted to see if student pretest measures were already affected by teacher measures, whether by teacher influence, or by non-random initial matching of teachers and students. We wanted to see if there were particular teaching approaches that worked better with some SES or ethnic groups than with others. Possible SES and ethnic differences on student pretests had to be identified, too. Main effects would be sought, but so would teacher-pupil interaction effects. Finally, ways had to be found or developed to synthesize findings across different outcomes, and to allow comparison of findings in the four replicated samples.

Our basic approach to the analysis has been a heirarchical multiple regression approach, with analysis of covariance, as the optimal way to assess change (see "D.P.H" et al., A Practical Guide to Measuring Project Impact on Student Achievement, 1975, p 54-58). All of the statistical procedures for carrying out various parts of such a system were already available, of course. What had never been done was to construct a rigorous, complete, yet cost-feasible system which would meet all of the objectives of our analytic plan. Simply to take a brute-force approach, putting all of the data through every necessary analysis, would have been prohibitively expensive in staff time and computer time. Thus, the final, nine-stage system took the form of a series of progressive screening techniques which only tested those relationships in a later stage that had proved significant in an earlier, logically prior stage. The final result is described hereafter as Program MANCHA, for Multivariate Analysis of Change.

This system is, in its way, a considerable advance in the state of the art for assessing change in any multivariate design which attempts to link outcomes to both entry variables and process variables. With suitable modifications, and respecification of terms, it could be used to assess the effects of psychotherapy, of alternative medical treatments, of experimental social progress -- indeed, treatments of any kind whose human and economic importance justifies this degree of effort.

#1 Student entering levels on achievement, coping skill, attitudes towards school and life, and self-esteem will all correlate positively with one another.

In order to design education that promotes student learning, without causing undesirable side-effects, it is necessary to evaluate several major outcomes simultaneously. Which outcome should one select to decide what kind of teaching is "most effective": Standardized test scores, effect on student self-esteem, or students' feelings about school? Just posing the question indicates how greatly oversimplified it is to look at only one effect of a year's instruction. We certainly want children to learn each year; but we also want them to keep wanting to learn in future years. Their feelings about school are just as important as their short-term learning, insofar as such feelings may influence their long-term learning. Stallings' study of Follow-Through programs (1975) found that drill and teacher pressure for results led to greater achievement test gains but poorer attendance; while "open, flexible" teaching had opposite, equally contradictory effects. Another study involved teachers whose successive classes over three years showed a consistent level of low, average, or high gain in standardized achievement test scores (Brophy, 1973). The teachers responded to the Adjective Self-Description, supplying self-reports of personal adequacy. The pupils evaluated the teacher's impact on them, including affective elements (SET). Those teachers whose pupils made the greatest standardized test gains also tended to be least liked by pupils for the way they made the pupils feel about themselves ($R = -.44$) and for the low interest value of their classes ($-.52$). Correlations between pupil gain scores and teacher self-evaluations showed a similar, negative relationship. For example, teacher's

"idealism" and "attractiveness" scores correlated -.56 and -.49 with pupil gains on the MAT. Teacher feelings about work, about parents, and about the future, on the Bown Self-report Inventory correlated -.30, -.53, and -.50 with pupil achievement gain. In short, in this study, pupil gain in achievement was accompanied by negative pupil attitudes about the way their teachers treated them. Furthermore, the higher the level of pupil gain, the less the teacher expressed self-esteem or a sense of taking control of his or her own life (Peck and Veldman, 1973). Finally, both Berliner (1975) and Borich (1976) emphasized the need for research that looked simultaneously at several kinds of outcomes.

The evidence cited in the review of the literature for this study (see Appendix D-1) indicates why these particular outcomes were chosen as critical elements in child growth, and therefore, critical goals for the educational process. Furthermore, the review indicates the interrelatedness among these characteristics. Specifically, the hypotheses are tested that self-esteem and achievement will make up a reciprocally reinforcing cycle; that achievement and attitude toward school will also form a reciprocally reinforcing cycle; and that coping skill will form a positive feedback loop with achievement and with self-esteem.

#2 Student Achievement Level, coping skill, attitudes and self-esteem will have reciprocal, positive effects on change in one another over the school year.

Where the first proposition proposed that there is a positive, dynamic interplay among the student characteristics at any given time they may be measured, this second proposition assumes that change over the course of a year in a given characteristic will also be positively affected by the student's initial status on the other selected characteristics. Thus, it is pre-

dicted that students who have low initial coping skills will gain less in achievement over the year than students who start out with higher levels of coping skill. Similarly, students with highly positive attitudes toward school are expected to gain more in achievement and in coping skill than those who have less positive attitudes toward school at the outset. Students with high initial self-esteem are expected to show greater achievement gain than those with low initial self-esteem. A reciprocal effect is expected in many of these combinations. Low initial achievement is expected to have a negative effect on changes in self-esteem; low achievement is expected to lead to smaller gains, or even losses in coping skill and in attitudes towards school; etc.

#3 Positive teacher characteristics will positively influence teaching behavior.

The presage measures of teacher characteristics such as the Adjective Self Description and Views of Teaching scales are expected to predict the COR and COS measures of teaching behavior during the year. Thus, ASD: Efficiency is expected to be positively related to COR-Systematic,organized and to COS: Time-on-task and Accuracy. The Views of Teaching: Attitude scale is expected to relate positively to the COR-KU and to the COS: Affect. Teacher experience is expected to be positively related to diagnostic and prescriptive skills based on the teacher interviews.

#4 Positive teacher characteristics will positively influence student behavior.

The teacher's age, experience, and personality characteristics are expected to show significant effects on student classroom behavior. For example, teacher ASD: Attitude is expected to show a positive relationship to student COS: Affect. Teacher experience and VOT: Coping are expected to show a positive effect on student COS: time-on-task and Accuracy. The mean SET score given a teacher early

in the year is expected to show a positive relationship to the COR-Pupil score for pupil behavior during the year; etc.

#5 Positive teacher characteristics will positively influence student outcomes.

Antecedent-consequent relationships are predicted such as the following. Teacher VOT coping skill will show a positive relationship to student outcome on coping skill and on achievement, measured over the year. Teacher ASD: Attitude is expected to affect student change in attitude toward teacher and toward school in general, over the year. Teacher VOT: Attitude is expected to have a positive effect on changes in student self-esteem. VOT: Coping is expected to have a positive effect on student coping and achievement over the year; etc.

#6 Positive teaching behavior will positively influence student behavior.

Whereas the preceding propositions concerned the beginning-of-year measures of teacher characteristics, this proposition concerns the process measures of teacher behavior in the classroom during the year and their predicted relationships to student behavior measures taken during the same period. For example, teacher COR-SO behavior is expected to show a positive effect on pupil COS: time-on-task behavior. Teacher COR-KU and COS: Affect are expected to show a positive effect on pupil COS: time-on-task and Affect; etc.

#7 Positive teaching behavior will positively affect student outcomes.

Specific hypotheses subsumed under this proposition are illustrated by the prediction that teacher COS: time-on-task behavior will positively affect student achievement and coping skill outcomes; that teacher COR-KU will show a positive effect on student attitude changes and changes in self-esteem; that teacher diagnostic and prescriptive skills will show a positive

effect on student coping skill and achievement outcomes; etc.

#8 Positive student behavior will positively affect student outcomes.

Sample hypotheses here include the following. Student COS: time-on-task will positively affect achievement and coping skill outcomes as well as self-esteem outcomes. Student COS: Affect will positively influence attitude changes over the year. Student accuracy will positively affect achievement gain. Student coping skill (SD) will positively affect coping skills and achievement changes over the year; etc.

#9 Student outcomes will often be affected, not just by teacher or student characteristics taken alone, but by their interacting effects.

It is more complex to state these hypotheses but they can be illustrated as follows. Students who begin the year high in coping skill are likely to be less strongly affected in their attitude and achievement changes by teacher Systematic-organized behavior than students low in initial coping skill. Students low in initial self-esteem are likely to be more vulnerable to negative effects of low Kindly-understanding behavior by teachers on their ultimate self-esteem than pupils with high initial levels of self-esteem; etc. The possible interacting effects of all teacher entry and teacher behavior measures with all student entry and student behavior measures will be tested.

#10 Each outcome will be produced by a unique set of factors. Some outcomes will be affected by a combination of SES, a teacher characteristic, one or two pupil characteristics and a class effect. Another outcome may be affected by sex, pupil attitude and teacher attitude, but not by demographic or class effects. Still other outcomes may be effected by ethnic or SES factors, operating to moderate the effects of one or more teacher or pupil characteristic.

Some outcomes may be explained entirely by a teacher main effect or by a pupil variable without any other necessary elements in the explanatory equation.

#11 Student socioeconomic status will be positively related to student entry measures.

A crucial, worldwide challenge is to invent educational strategies that will reduce the very large learning gap between middle-class and working-class children. In a ten-year, cross-national study, "Coping Styles and Achievement," carried out in eight countries (Brazil, England, Germany, Italy, Japan, Mexico, the United States, and Yugoslavia) the most outstanding finding is not one of national or sex or age differences, but the large, systematic deficit in all countries in educational achievement of skilled working-class youth compared with upper-middle-class youth at ten and fourteen years of age (Peck, et. al. 1972; Peck, et. al. 1973). What is more, the working-class youth in most of the countries show many parallel deficits in coping skills and in self-esteem. As might be expected, numerous interaction effects are observable with such factors as class, age, and sex.

The socioeconomic differences are actually underestimates of the true discrepancy. For example, at least fifteen to twenty percent of the Anglo ten-year-old, working class children in school in Texas proved to be functionally illiterate and had to be excluded from the sample used in this cross-national study. Furthermore, many of the academically least effective students in most countries had dropped out of school by age fourteern.

Since it is all too well established that children from the lowest socioeconomic status (SES) level (who were not included in this study) show even poorer academic achievement in all countries, the conclusion is inescapable that none of these countries has found an effective way to educate the children from the lower sixty to seventy-five percent of its population. This finding

parallels the findings from the International Study of Educational Achievement (Husen, et al. 1967; Husen, 1972), the Coleman report (1966), Jencks, et al. (1972), and many other studies (e.g., Bryant, et al. 1974). Standard schooling works very differently for these children than for middle-class children. In the United States, this means that the children from the lower half of the socioeconomic range (at least sixty-five percent of the elementary school population) are much less likely to master the standard academic skills than are their middle-class schoolmates. There is, in short, a powerful "interaction effect" whereby the absence of a differentiated treatment produces this large inequality of educational outcomes.

Tyler (1951) proposed that this deficit might be made up by providing lower-class children with some of the practice that middle-class children get outside of school. Merely adding more hours to the school day, with the same instructional procedures, probably would not get to the heart of the problem. Brophy and Evertson (1975) report systematic differences in teaching techniques that produce achievement-test gain with young students of low and high socioeconomic status. Stallings' report (1975) contains perhaps the largest, most solid body of evidence ever assembled which bears on this issue.

Perhaps another, crucial key to the problem comes from the evidence of important social class differences in parental example (Marjoribanks 1971) and in parental "teaching techniques" (Blau, 1972; Greenberg and Davidson, 1972; Kohn 1969). A number of pilot programs have been mounted in the past decade by school systems and by state education agencies for the purpose of innovation. Some of them reach out to involve lower-class parents much more actively and continuously in the schooling process; some, to provide new kinds of training for parents which help them teach their children coping skills that give them a better preparation for school. All of this together, if it were tested, systematized, and made a permanent part of our educational system (in the

broadest sense of that term), might substantially improve the motivation, the coping skills, and the academic achievement of our undereducated majority.

The hypotheses will be tested that student SES is positively related to student level on each of the entry measures; in particular, achievement, coping skill, and self-esteem, to judge by previous research evidence.

#12 Student socioeconomic level will positively affect student outcomes.

This tests the idea that schooling has the unfortunate effect of widening the gap, each year, between high status and low status students, not only in achievement, but in the attitude,self-esteem, and coping outcomes. It is expected that SES will sometimes have a direct, main effect of its own but more often will show an interacting effect with a teacher or student variable. In the later case, high SES students might be predicted to be less negatively affected in their self-esteem by teachers low on COR-KU than low SES students; etc.

#13 Ethnic differences will be found in the student entry measures.

Almost all of the evidence to date suggests that Anglo students are likely to have higher mean scores in achievement and possibly in coping skill and self-esteem, than Black or Chicano children. The purpose of this inquiry into ethnic effects is certainly not to bolster unfortunate stereotypes which can have unfair and unfortunate effects on minority children. To the contrary, the purpose of this inquiry is to see if it is possible to identify teaching strategies that may be differentially more effective for different ethnic groups. The identification of beginning differences is merely a first step in that inquiry process.

#14 Ethnic differences will affect student outcomes.

This is the place where the most important findings are sought. Although possible main effects of ethnicity on various outcomes will be tested, the crucial inquiry is to look for interaction effects with other pupil and teacher variables. For example, it may turn out that Chicano children gain more in self-esteem when placed with teachers who are very high on Systematic-organized behavior whereas Anglo children gain more with teachers who are high on Kindly-understanding behavior. Black children who begin the year low in self-esteem might make much smaller achievement gains than Black children who are high on self-esteem, whereas this factor of self-esteem might not be so critical in the Chicano or Anglo groups. It is by looking to such specific patterns that it may be possible to find out what will work best and worst for the children of each ethnic group, to the degree that there may be differences in their reactions to instruction.

#15 A substantial proportion of the significant relationships among the variables tested will prove to be curvilinear.

The quadratic form of all predictor variables will be included in all of the analyses, until a particular predictor proves to show no curvilinear relationship to the other variables concerned in further analyses. As the review of the literature (Appendix D-1) demonstrates, extreme levels of a characteristic, either positive or negative, can have a different effect on another variable than a medium amount of the characteristic. For example, a moderate amount of Systematic-organized teaching behavior may prove to be considerably better than a minimal amount in its effect on student learning or attitude; but a high degree of systematic behavior might prove to have a less beneficial effect, particularly with certain kinds of students, than a moderate amount of organi-

zation. Another example might be that pupil achievement would not be substantially affected by self-esteem at the middle or high levels, but might be adversely affected if self-esteem were very low. The analytic system will first test for the significant operation of a quadratic form of each predictor variable before deciding whether to retain that form in each subsequent analysis. This insures that no curvilinear relationships will be overlooked.

#16 Each student instrument will cross-validate in up to seven ways.

The details of the cross-validation process are described below in the Analysis section and in the Validity sections of each instrument manual contained in Appendix A. The selection and arrangement of instruments shown above in Figures 2 and 3 were made with this purpose in mind. Any one process measure of student behavior, for example, is checked against at least one other source of information (e.g. COS observer codes compared against teacher ratings on the Student Description). The process measure is checked against those presage measures which should logically predict it. The process measure is also checked against the outcome measures which it should logically predict. Specific examples have already been cited under the earlier propositions in this list.

#17 Teaching behavior will be validly measurable.

If a considerable part of teaching behavior is an expression of the teacher's personality and values, and if individual personality is persistent across time, it should be possible to demonstrate individual consistency in teaching behavior if the right kinds of measures are taken, at the right levels of generality. This would seem to argue for fairly high inference measures which allow for different "phenotypic" expressions of stable "genotypic" traits.

Different classroom situations may call for different "phenotypic" actions, such as low inference coding systems address.

The evidence of earlier studies has been equivocal; but a substantial and increasing amount of positive evidence indicates that some important dimensions of individual teaching behavior do tend to be consistent across time and contexts. Shavelson and Dempsey-Atwood (1976), in their review of existing stability data, concluded that not much consistency of teacher behavior has been shown, although Brophy (1976) pointed out many methodological limitations in the studies they cited. A longitudinal study of over 1,000 elementary teachers, however, reports "a marked to very high correlation (.42 for college advisor-pairs; .80 for first year by third year evaluator) between the six evaluators of the teacher from student teaching to the third year in the profession." (Myers, 1967) The evaluative scales have to be examined in more detail, of course, to see what kinds of teaching behavior were thus evaluated.

Sears (1966), Simon (1967), and Pfeiffer (1967) also found considerable consistency in teachers' behavior from year to year, or across classes. Some other studies, using low inference measures, found more variation, often attributable to the constraints of different subject-matter and different classroom structures (Marshall, et.al, 1976). Brophy and Evertson (1976), however, found consistency coefficients of .49 to .74, and higher, on numerous low-inference variables when junior high school teachers were compared across two different classes. High-inference ratings on the Emmer-Peck scales, based on the same observations, ranged from .73 to .86. An earlier study (Brophy, et.al, 1975) found high teacher stability, for most of these scales, across two school years.

In designing the current Teaching-Learning Interaction Study, provision was made to test the proposition of teacher consistency in several cross-checking ways. First, teaching behavior was rated with Ryans' Classroom Observation

Record: twenty scales that reproduced Ryans' factors of Systematic-organized (SO), Kindly-understanding (KU), and Stimulating-imaginative (SI) behaviors, in social studies classes, chiefly.

Many of the same teachers were observed again in the following year in a new subject (reading), with new classes of students. A team of half-old, half-new observers achieved the same level of independent rating reliability as the first-year team.

In each year, the observers also rated student behavior on eight scales, yielding a class factor score of "Positive Work Orientation." This could be considered, for the present purpose, as a Teacher-product measure, to be tested as a presumed "consequence" of the teaching behavior measures. (See the Validity sections of these instrument manuals in Appendix A.)

An independent appraisal of the teachers by their students was carried out in each year, using the eight scale Student Evaluation of Teaching (SET) form. This produced an "Effective Teaching" factor score for each teacher in each year, derived from different classes of students. In Year Two, there was also a direct measure, by student report, of the teacher's impact on their self-esteem, How This Class Makes Me Feel. This can be treated as a consequent measure, and compared with its logical antecedents (COR, e.g.).

In the second year, in particular, the observers also recorded highly specific kinds of teaching behavior using the COS behavior codes. Finally, the same staff members, acting as teacher interviewers, made ratings of the teachers' diagnostic and prescriptive skills, and follow through behavior, based on the interviews.

Finally, the self-report instruments completed by the teachers, the ASD, VOT and Concerns Checklist, provided measures of many of the same constructs which were assessed by the pupil and the observer measures.

Cross comparing these five kinds of measures from three independent sources, if the simultaneous measures agree with one another, this cross validation would tend to confirm the proposition that teaching behavior is validly measurable. This issue is centrally crucial to the entire study, of course, since unless teaching behavior can be validly measured, there is no point in undertaking research on the effects of some unknown quantity known as "teaching".

#18 Individual teaching behavior will prove relatively stable across different occasions and different situations when it is defined at the proper level of generality.

The COR observer ratings are taken on repeated occasions, as are the COS behavior codes and the Interviewer Perception ratings. The ratings are first tested, of course, for a satisfactory,high level of rater reliability before moving to the measurement of reliability of ratings across different occasions.

One of the most important reasons for the replication feature of the research design was the opportunity to compare the measure of teaching behavior taken on the same teachers in two different years, with different classes of students, and usually teaching different subjects. In Austin, for example, complete data on all measures were available in both years for a subset of twenty-eight teachers. In Kentucky, complete data were available in both years for a subset of twenty-two teachers. (The total number of teachers in Austin who were present in both Year 1 and Year 2 was 33, but not all of the data was available on every one of these teachers.) Thus, the COR measures for a given group of teachers could be correlated across the two years to determine how similarly the observers saw the teachers acting in the two years. The SET rating of the teacher by two different classes of pupils in the two years

could similarly be compared. The interviewer ratings in both years could likewise be compared. Finally, in a more complex way, the idea of the stability of teaching behavior could be inferentially tested by seeing if the patterns of relationship between the teaching behavior measures and the "teacher-product" measures (COR: pupil and How This Class Makes Me Feel) compared across the two years.

These procedures actually constitute a quite stringent test of the proposition since both the students and the subjects taught varied across the two years. Secular changes in the teachers lives, changes in school administration or other contextual factors, and other unknown changes that might have occurred from one year to the next, all would tend to reduce the amount of similarity that could be found in behavior across occasions and situations. If substantial consistency were found, it would seem to indicate very genuine stability of individual teaching behavior.

#19 Correction for instrument reliability will improve the accuracy and validity of findings.

As is discussed in several places in this report, we have discovered during the study that the state of the art is not yet ready to allow us to test this proposition with any certainty. The correction procedures which have previously been developed either increase error or are inapplicable to a design with two or more input variables. It remains to be seen whether Joreskog's method is applicable and practically feasible for this kind of multi-variate design. That will be pursued in the coming year.

#20 Findings will replicate across samples.

Actually, we anticipated that some findings would not replicate, not only

because of chance effects, but because real differences in the samples of students and in their reactions could lead to a genuine difference in the relationship patterns found in the two different years and in the two different places. Differences across the sites could also be real because of differences in the teacher samples which were not accounted for by the measures assessed in this study.

Indeed, the replication test as a measure of the validity and generalizability of findings is an extremely stringent one. It is not just possible but certain that a great many contextual factors, and a great many intrapersonal factors,operate to influence student changes, over and beyond the variables we have been able to identify and measure in this particular design. All such factors would tend to reduce greatly the chance of repeated findings, not because there is a lack of lawfulness in human behavior but because the outcomes depend on a much larger set of factors than any single study has been able to encompass including this one.

Nonetheless, the inclusion of repeated samples is a centrally important feature of the research design. Although commonly used in the biological and physical sciences, and frequently applauded in theory for any scientific study, replication has rarely proved affordable in the behavioral sciences. To the best of our knowledge, there has never been anything close to an accurate replication of any study as complex as the present one. Rather than hope for the implausible, the replication by other researchers of a study this complex, difficult, and expensive, which is not of their own design, we decided it was essential to provide for replication within the design of this study itself.

Even so, the replication has not been perfect in all details. Some changes had to be made in instrument construction or selection, when it was

found that the research samples were hitting the ceiling on a pretest, for instance, or where some items were proving too objectionable to teachers or students. One instrument attempted in the first year was abandoned because a substantial proportion of the teachers would not accept it. (Directed Imagination). The observation system was substantially expanded and differentiated in the second year after the first year's experience taught us how much the observers could cover. In such places in the data, special analyses of subsets of items will have to be done, using just those items common to both years, as a check on the comparability of the total instruments in their two different versions. Similar checks on the comparability have to be made on instruments where other kinds of differences or additions occurred.

For the most part, however, the main body of data in both years is directly comparable. One of the procedures most vulnerable to local variation, the observation by different teams in two geographically separated sites, was met by having a common training of all observers,to a high degree of reliability, before they separated to operate in the two different parts of the country. After that, a check for "drift" was necessary at a later time in order to maintain a good degree of comparability of the observation-based scores. The chief guard against unwarranted assumption of similarity, however, is the independent analysis of the data in each of the two sites, with the comparision coming at the point of the findings, not through a pooling of the data from both sites before the analysis is done.

INSTRUMENTATION

Instrument Search

In January 1974, the R&D Center's Search and Analysis staff began a search for machine-scorable self-report instruments that would measure some selected teacher and student variables that would be likely to affect designated student outcomes. Component variables were sorted into the cells of a 3 x 3 grid of skills (cognitive, affective, and coping, following Bloom, 1956, Krathwohl, 1964 and Peck, 1967) and behavioral domains (interpersonal, intrapersonal, and career-related knowledge). (Figure 1) The basic criteria for selecting instruments were content area, machine scorability, adequate reported validity and reliability coefficients, and appropriateness for administration to sixth graders.

Such instruments are not the ideal way to assess the individual characteristics that the study proposed to investigate. Behavioral measures, including depth-interview material, would be far more desirable for assessing coping skills, attitudes, self-esteem, and even academic skills. Self-report measures have quite imperfect validity for estimating these actual characteristics. Free-response instruments display more of the individual-at-work than do forced-choice instruments. The sole justification for choosing self-report, machine-scorable instruments was that of cost. Not only was this the only affordable approach, with a sample large enough to sustain the multivariate design of the study; it is the only approach to a mass-assessment system that a school district might be willing to finance, if it proved to have some useful properties. Under these circumstances, whatever relationships the measures might demonstrate in this study will probably be underestimates of the true relationships among the variables which the instruments are supposed to represent.

The search and analysis process for identifying usable instruments was arranged in a cost-efficient series of increasingly fine screening procedures. After familiarizing themselves with the background of the project and the contents of the cells of the conceptual grid mentioned above, staff members developed an instrument evaluation sheet to indicate quickly critical features of a given instrument and to expedite the writing of instrument descriptions. The evaluation sheet included the following items: (1) the name of the instrument; (2) author('s') name(s); (3) cost of the instrument; (4) length of the test and (5) time required to administer the test; (6) the address of the disseminators of the test; (7) date of copyright; (8) explanation of scoring; (9) test reliability; (10) validity coefficients; and (11) a general description of the instrument.

Computerized searches were conducted to locate references to instruments that are indexed in the ERIC CIJE (Current Index to Journals in Education) and the ERIC RIE (Research in Education) systems. As documents that included instruments or references to instruments were located, reference lists of those documents were systematically monitored, often leading researchers to additional sources. In addition, all books pertaining to instruments in The University of Texas system libraries were searched. Two UCLA sources were also carefully combed by the project director: CSE-RBS Test Evaluations: Tests of Higher-Order Cognitive, Affective and Interpersonal Skills (Hoepfner et al., 1972) and Elementary School Test Evaluation (Hoepfner et al., 1970).

Four hundred nine instruments were surveyed in this extensive search. Of the 409, only 55 (13.4%) received high ratings for both reliability and validity. Thirty-nine instruments received a high rating for either reliability or validity; 284 instruments were considered to be low in both validity and reliability. For the remaining 31 instruments, information was incomplete and

unavailable, and they were not classified. Listings and classifications of instruments are appended to this report.

Instruments and Variables

The general schedule for administering the student instruments in the Teaching-Learning Interaction Study specified that measures be gathered in fall (pretest) and in spring (posttest). The teacher instruments were collected only once (except for retest reliability samples). Classroom observations were carried on throughout the school year. The posttest period for test-retest reliability computations followed the fall pretest administrations by three weeks.

The instruments used in the first year of data collection, and the variables they yield, will be discussed in this order: demographic variables, student attitude and opinion instruments, student achievement instruments, teacher demographic variables, teacher attitude instruments, classroom observation measures, and teacher interview systems. The instruments and variables of the second year of data collection will be introduced in a similar order; modifications of Year 1 instruments for use in Year 2, and the addition of new measures, will be noted in the text (see Appendix A for complete instrument manuals).

Year 1 , Student Measures

Student demographic data are contained in the Pupil Identification Number and the Student Biographical Form. The Pupil Identification Number codes the following information: year of study, school attended, grade, class/teacher, sex and ethnicity. Information carried in the Identification Number was reported for the pupils by teachers and project staff members. The Biographical Form, designed by R. F. Peck, garners the following information: child's name, age and birthdate, number of older and younger siblings, father's occupation and education, and mother's occupation and education. Family socioeconomic status (SES) is calculated according to a system developed by Robert J.

Havighurst for the Coping Styles and Achievement: A Cross-National Study of School Children research (Peck et al., 1972).

Data on students' self-concept were collected using the Piers-Harris Children's Self-Concept Scale (P-H), developed by E. Y. Piers and E. D. Harris (Piers, 1969). In pretest and posttest administrations students noted the applicability or truth of 80 statements about themselves. While six factors may be scored separately, for the Teaching-Learning Interaction Study only the total score is used. The internal consistency reliability of this score was .93 for both pre- and post- measures among the Austin sixth graders. Publisher-reported test-retest reliability ranges from .71 to .77 (t = 4 mos.).

Students' attitudes about the school situation and their teachers were assessed with the School Sentiment Index and the Student Evaluation of Teaching instruments. The School Sentiment Index (SSI) was developed by Sandra Frith and Diane Narikawa as part of the Attitude Toward School battery distributed by the Instructional Objectives Exchange at the University of California at Los Angeles (IOX, 1972). In its published form, the SSI consists of 81 statements that represent five dimensions: teachers, learning, social structure and climate, peers in school, and school in general. Students indicate whether each attitude item is true or false for them. The form administered in fall and spring of the 1974-75 school year consisted of 54 items listed on three stapled sheets of paper. The "school social structure and climate" and "peer" dimensions were omitted, and some single items that were deemed abrasively negative toward teachers were reworded or omitted. Scores entered on the final data file are: teacher mode of instruction, teacher authority and control, teacher interpersonal relationships (these subscales make up the teacher dimension), and teacher total; student attitude toward learning; student attitude toward school; and SSI total.

SSI total is the score that is used in the basic analyses of data for this study. The internal consistency reliability of this total score, based on measures of our first year sample of children in Austin was .92 (pretest) and .94 (posttest). Test-retest reliability (N =) was .

The Student Evaluation of Teaching instrument was developed through successive versions by Donald J. Veldman and Robert F. Peck of the UTR&D Center from 1959 through 1970. It was modified for use in the TLI study by Peck and Elaine Morris. A single score of Attitude Toward Teacher is derived from students' responses that indicate agreement or disagreement with ten Liker-scaled phrases that describe the teacher. Internal consistency reliability coefficients (alpha), based on fall and spring administrations of the SET to Austin sixth graders, were .76 (pretest, N = 1333) and .80 (posttest, N = 1297). Cross-class validity in an earlier study of 50 teachers averaged about .85 (Veldman, 1970).

Measures of students' general attitudes and coping skill were provided by the Student Sentence Completion (SSC) form. No other satisfactory measure of coping skill was located in the instrument search. Free-response forms of the SSC were developed during the 1950's by Robert F. Peck. Peck, Elaine Michelis and many collaborators in the Cross-National Study of Coping Style and Achievement (Peck et al., 1972) developed a modified free-response form. Peck used the extremely reliable coding and scaling systems and the validity findings from the Coping Styles study to develop a multiple-choice instrument for use in the Teaching-Learning Interaction Study in 1974. The SSC was scored for 26 of 29 sentence stems and response-alternative sets after item analyses showed that one item did not correlate significantly with the appropriate scale, and that two items whose wordings were modified between the first and second years of the study did not have the same meaning across the two years. Two scales,

Attitude and Coping, are scored from two separate sets of items. Pre and post scores for each scale are entered in the student data merge file. For Austin sixth graders in the 1974-75 school year, SSC-A internal consistency reliability (alpha) was .61. For SSC-C, internal consistency reliability was .63. Test-retest reliability coefficients (t = 3 weeks, N = 90) were .63 and .68 for the Attitude and Coping scales, respectively.

A peer rating of students' school-related coping skill was employed for the Teaching-Learning Interaction Study. The Behavior Rating Scale was developed by Robert Peck in 1965 for the Coping Styles and Achievement Study. Ratings are obtained by asking students to rate nine of their peers and themselves. To insure that each student will rate and be rated by nine others rather than an entire class, an economical sampling procedure known as the Peer Jury Rating System is employed. In the first year of data collection, the BRS was used only once, in fall 1974. The BRS consists of 13 bipolar items, single words or phrases separated by a five-space, Likert-type response scale. Two items were deleted from scoring because of low item-total correlations. Mean item scores for each individual are computed from the peer ratings and entered separately as BRS-SR (self-rated). Internal consistency (alpha) of the 11 item scale was .95 for peer ratings (N = 1379). Test-retest reliability (N - 90 subjects with 810 peer ratings) is for Austin Year 1 and for Austin Year 2 (N = 60 subjects with 540 peer ratings).

In the first year of data collection for the Teaching-Learning Interaction Study, students' academic achievement was measured with the Comprehensive Tests of Basic Skills, developed by CTB/McGraw Hill of Monterey, California and published in 1973. This was the collective preference of the collaborating teachers and the local and national consultants on testing who advised the project. CTBS forms are available to measure students' skills in the language arts,

social studies, arithmetic and science. There are several forms of the tests and several levels of difficulty. Procedures for comparing scores on different forms and levels of the tests are provided by the publisher (CTB/McGraw-Hill, 1974). For pretesting students in Austin, Texas and Daviess County, Kentucky CTBS Form S, Level 2 was used. Austin students were tested either in language arts (language mechanics, language expression) or in social studies, depending upon the classes in which they were to be observed. Kentucky students were tested in language arts, social studies, arithmetic, or science, depending on the class in which they were observed. In Austin, all students completed CTBS, Form S, Level 3 in their respective subject area for the first year posttest.

Level 3 was given as the posttest because more than 25% of the students scored at or near the ceiling of the Level 2 form, in the fall. In order to achieve comparability, scores from Level 2 and Level 3 were converted to extended standard scores, following the publisher's procedure. Unfortunately, it appears that this "equivalence" scale may not accurately equate the two forms of the test, for the total Austin sample actually scored lower on this scale in the spring than they did in the fall. While this test, like all of the other widely-used standardized tests (which had been examined in detail), bears only the slightest, most generalized relation to the specific curricula content actually studied in these classrooms, it seems most unlikely that the students were really more ignorant in the spring than they had been in the fall. Examination indicates that the Level 3 test is a good deal more difficult than Level 2. Our guess is that the publisher's conversion tables somehow do not fully neutralize the effect of this difference, so that an apparent "loss" can appear when the two forms are used pre and post, as was done here.

In Kentucky, with a few exceptions, the first year posttest was CTBS, Form S, Level 3. The exceptions were some fourth and fifth grade classes who were posttested with CTBS Form S, Level 2 and some sixth and seventh graders who were pretested in social studies using CTBS, Form S, Level 2 but were given CTBS Form Q, Level 3 in language arts for their posttest in spring 1975. Nine Kentucky classes did not have both pretest and posttest in the same content area and must, therefore, be omitted from the Achievement Analyses of first year data for the Teaching-Learning Interaction Study.

Internal consistency reliability coefficients and interlevel correlation coefficients are reported for CTBS (CTBS, Technical Bulletin No. 1, 1974, p. 41,42). Form S, Level 2 Language Arts test has reliability (KR20) of .82 (language mechanics) and .99 (language expression) for students in late fifth grade or early sixth. Form S, Level 3 (the posttest) has reliability of .76 (mechanics) and .87 (expression) for late sixth grade administration. Interlevel correlations for those scales on the language arts test are: $r = .68$ (mechanics) and $r = .77$ (expression) for sixth graders. Form S, Level 2 Social Studies test has internal consistency of .87 for late fifth, early sixth graders. Form S, Level 3 Social Studies (the posttest) has internal consistency of .89 for late sixth grade administration. Interlevel correlation of those tests is published as $r = .73$.

For fourth graders, the Language Arts test, Form S, Level 2, has internal consistency reliability of about .93; for fifth graders internal consistency reliability is about .93; for seventh graders it is .93 at the pretest administration. For each grade sample that was posttested with the same level test, internal consistency reliability rises slightly, by .01.

Internal consistency reliability of the mathematics test, CTBS, Form S, Level 2, for fourth graders is about .94, for fifth graders about .95, for sixth graders about .96. Again, for each grade sample posttested with the same level

test as at pretest (fourth and fifth graders), internal consistency rises slightly, across the year. Sixth graders posttested with level 3 after pretest with level 2 in mathematics should evince internal consistency of about .95; seventh graders' test internal consistency reliability on level 3 math should be about .96. Interlevel correlation on the math test is r = .67 for sixth graders; no correlation is published for seventh graders.

For the science test, Form S, Level 2, internal consistency reliability is about .79 for fourth graders, .83 for fifth graders, and .86 for sixth graders (seventh grade reliabilities are not reported for level 2). Students posttested with level 2 will have slightly higher internal consistency reliability coefficients. Sixth and seventh graders' internal consistency on the science test, level 3 are reported as .86 and .87, respectively. For sixth graders, level 2 with level 3 interlevel correlation r = .75.

For the social studies test, Form S, Level 2, internal consistency reliability ranges from about .84 for fourth graders, .87 for fifth graders, to .89 for sixth graders, level 3 reliability is .89 for sixth graders and .91 for seventh graders, at the end of the school year. Interlevel correlation of level 2 with level 3, social studies test, is computed as r = .73 for sixth graders.

CTBS test norms are based on the results of administrations to a large stratified random sample of school children in 1972 and 1973. For CTBS Form S, Level 2, 3,673 fourth graders, 9013 fifth graders and 5,116 sixth graders were tested. For CTBS, Form S, Level 3, 4,688 sixth graders and 13,922 seventh graders were tested.

Year 1, Teacher Measures

As they entered the project, teachers filled out a 23-item Biographical Form. The instrument consists of 19 items that ask the teacher to provide

information about her/his education, experience, home and family. In addition, a self-assessment section of four questions requests the teachers to write short answers regarding what they consider their greatest personal strengths and how they relate to teaching effectiveness, what teaching skills or techniques they would like to improve, what the most important thing in their life is, and what plans they have for the future. Teacher ethnicity, sex, age, academic degree, and years of teaching experience are coded from this form and entered in the data merge files for use in the project analyses. Data are available from which teachers' "family of origin" SES may be derived. Although it is not used in the basic analyses, this figure may be fruitfully utilized in assessing the effects of background on teacher characteristics.

Additional self-report information about teachers was gathered with the Views of Teaching (VOT) instrument, a sentence completion form developed by Robert Peck. As teachers entered the project (most of them in fall 1974, at the project's inception) they were asked to complete the form.

VOT is a free-response adaptation of the Student Sentence Completion; 14 of the 38 stems are identical, and both Attitude and Coping scales are scored. Each scale measures behavior across items dealing with authority, interpersonal relations, task achievement, anxiety, handling children in the classroom, and self-concept. The variables are scored with a modification of the original system developed in the Coping Style Study. VOT-Attitude and VOT-Coping show internal consistency reliability (alpha) of and .69, respectively, for 56 Austin sixth grade teachers. Test-retest reliability, based on a subset of 32 Austin teachers, is .53 for the Attitude scale and .73 for the Coping scale.

Teachers also completed the Adjective Self-Description (ASD), an instrument developed by Donald J. Veldman and G.V.C. Parker at the UTR&D Center. ASD

consists of 56 Likert-scaled adjectival items; the subject is asked to rate each descriptive item on a scale of one to five according to how well it describes him or her. Seven factorially distinct scales are scored and entered as discrete variables on the data merge files for use in analyses of project data. The scales are Attitude, Behavior, Efficiency, Introversion, Anxiety, Ideology/individualism and Attractiveness. Veldman (1971) reports internal consistency reliabilities (alpha) for the seven scales: .85, .75, .84, .88, .83, .64 and .69 respectively. Internal consistency reliability coefficients for Austin teachers are: __ , , , , , , ; for Kentucky teachers: , , , , , , .

A story-eliciting, self-report instrument, the Directed Imagination (DI) was administered to many of the teachers in fall 1974. This instrument was developed at the UTR&D Center. Subjects are instructed to write four short fictional stories about teachers and their experiences. Fifteen variables or scales are scored for the DI. Internal consistency reliability ranges from .53 to .82, based on a sample of 250 protocols provided by 125 female education majors. Four validity studies tend to confirm the predictive and construct validity of the Directed Imagination instrument (Veldman and Menaker, 1968). Analysis of scores derived from the protocols provided by the TLIS teacher sample, however, led to the decision that the DI would not yield meaningful measures for this study. The most significant problem arose from the fact that a sizable number of teachers either refused or "forgot" to complete the instrument. This was the only instrument that was negatively received by the teachers. While DI scores will not be used in the statistical analysis of TLIS data, some information from the instrument is included in the naturalistic descriptions of teachers.

At the time of posttesting in spring 1975 teachers completed the Student Description (SD) instrument with regard to the six or seven Special Study students in their classes. SD, designed by R.F. Peck for use in the TLIS, is composed of the school adjustment items of the BRS and the attitude and behavior items of Ryans' Classroom Observation Record-Pupil factor (COR-P). The teacher was asked to rate how each Special Study student usually acted; this yields the SD-Status score.

In spring 1975 teachers completed the School Preference Questionnaire, developed by R.F. Peck and Elaine Morris during 1975. The Questionnaire consists of three parts: a section in which the teacher checks descriptive statements applicable to the current teaching situation, mostly about teaching load, grouping and planning; a bipolar, Likert-type response format in which the teacher may make known his or her preferences about teaching and administrative responsibilities; and the Class Checklist section of 10 phrases that the teacher applied to the homeroom group in comparison with all previous groups in that teacher's experience.

Austin teachers also completed the Teacher Needs Survey, included in their instrument package as a service to the Austin Independent School District. The survey asked teachers to note whether greater access to 20 specific materials, services or facilities would improve their teaching effectiveness or make the school environment more comfortable. Additional comments were invited. Teachers were told that their responses would be tallied and forwarded (anonymously) to the appropriate administrators, as was done.

Year 1, Classroom Observations and Teacher Interviews

Between January and March, 1975 each classroom was observed every two weeks and the teachers were interviewed once about individual Special Study students' learning. The team of observers used three separate forms to gather information: The Classroom Observation Record, adapted from Ryans (1960), with which the

observer rated teacher-student interactions using 24 items; the Classroom Observation System, a set of coding categories that describe the observed behavior of the teacher and each of six or seven students, at regular intervals during the class period; and the Classroom Observation Report Form, a descriptive report of the observed interactions which illustrates the codes and ratings given to the teacher and the student. The observer completed the three instruments during and immediately following a 20-minute observation session in the classroom.

The Classroom Observation Record (COR) was developed by D. G. Ryans and published in Characteristics of Teachers (Ryans, 1970). The COR is used by a trained observer to rate the behavior of the classroom teacher and the class as a whole. The form consists of two sets of bipolar scales, one set describing the teacher and one set describing students. Specific behavioral definitions are provided to aid the observer.

The form used first in the TLIS consisted of eight items regarding pupil behavior and twenty items regarding teacher behavior. Four student items were added to Ryans' original instrument. They are "unhappy-happy;" "low time-on-task/high time-on-task;" "low cognitive level - high cognitive level;" and "chaotic-controlled." Three new teacher items, "low cognitive lecture - high cognitive lecture;" "low cognitive questions - high cognitive questions;" and "unclear - clear" were added. One item appearing on Ryans' original instrument, "immature - integrated," was rephrased "self-centered/student centered" to better convey the meaning of the scale to classroom teachers. Ryans' item, "unimpressive - attractive" was dropped after pilot work indicated that, for a subsample of Austin teachers, it was not significantly related to any of the main factors.

Factor analysis of the means of the 20 scale scores for Austin teachers (N = 53) indicated that three independent factors were measured by the instrument. The three factors are almost identical with those identified in the earlier study by Ryans, and were given the names: Kindly, understanding (KU), Systematic, organized (SO) and Stimulating, inventive (SI). Internal consistency reliability for the three factors (unit weighted) ranged from .97 to .98, based on 2 or 3 observations of 53 teachers by 4 observers.

While both Austin and Kentucky Year 2 samples confirmed the three designated factors, the Year 1 Kentucky sample did not. Only a two-factor solution was obtained from the COR ratings of those 27 teachers. Those two factors correspond to the SO and KU factors of the three factor solution. It has been suggested, largely on the basis of observer reports, that the Kentucky teachers manifested considerably more highly "Stimulating, inventive" behavior than the Austin sample, and that the absence of a range of behavior on those SI factor items would collapse the factor. The judgment was made to score the COR for the three factors on all teacher samples, recognizing that the SI factor for the Kentucky Year 1 sample probably would merely replicate the patterns shown by KU and SO.

The Classroom Observation System (COS) was designed by R. F. Peck and the project staff at the Research and Development Center for Teacher Education to meet particular objectives of the TLI project. It includes ideas from Holley, Bowden and others of the Austin Independent School District, from earlier work at the R&D Center, and from the work of Stallings, Flanders, Moskowitz, O'Leary, and Berliner. The system includes time-on-task and time-off-task codes to incorporate Bloom's suggestion that time spent on-task is strongly and positively correlated with learning (recommended to us by Garry McDaniels).

Every minute for 20 minutes the observer observed and recorded three aspects of the classroom situation, in addition to recording descriptive notes.

The three aspects of the classroom situation were the classroom activity, the cognitive level of the activity and the on- or off-task behavior of the selected Special Study students in the classroom. The activity and cognitive level were observed and coded at the beginning of the minute. The Special Study students were each observed for ten seconds and then their task behavior was coded. Observers also made short notes describing the pupils' behavior, coping skills, and affective state, any interactions between the teacher and the Special Study student, and the teacher's own teaching methods or coping styles during the classroom activity.

The Classroom Observation Report Form (CORF) was completed when the observer left the classroom following an observation session. The CORF was designed by R.F. Peck and the project staff to record observer impressions about the classroom (noise level, lighting, etc.), teacher and student behavior in general, and critical incident anecdotes concerning the Special Study students. CORF information is used in constructing case-study reports on teachers and students, to illustrate the behavior represented by the statistical findings.

During the first year of data collection teachers were interviewed once, around the time of posttesting on the student measures. Teachers were interviewed by a TLIS staff member who had observed in their classes.The interview was structured by the Teacher Interview Guide: Individual Learning Needs form. Questions focused on teacher knowledge of, and response to, the learning needs of each of the Special Study students. The teacher was asked what he or she thought each student had needed to learn during the school year and what had been done to meet these needs. The teacher's methods of assessing the student's progress, materials and resources that proved helpful, the student's responses to them, and the results of end-of-year assessment were all addressed. The information gathered in this interviewing session is also used in writing the descriptive case study reports on teachers who participated in the TLIS.

Year 2, Student Measures

There were few changes in instrumentation for the second year data collection. One new student instrument was added as a posttest in Year 2, and the measures of achievement were changed. The demographic information that was gathered was altered only slightly. The Student Identification numbers for Year 2 follow the same format as in Year 1. The Student Biographical form was altered slightly. The original version of the form asked for the number of older brothers and sisters and the number of younger brothers and sisters; the revised version asks for the name and age of each older brother,younger brother, older sister and younger sister. An additional sibling-related section asks for the name of the twin brother or sister if the subject child has one. The revision of sibling information questions was effected to provide data on birth order. This information is not used in the main analysis of TLIS data but may be used in subsequent analyses. The additional questions increased the length of the form from one to two pages.

The Piers-Harris Children's Self-Concept scale, the measure of self-esteem, was revised and renamed the Student Self-Description. The revision was the reduction from 80 to 33 of the number of items. Several sentences which contributed to each factor were retained, while the marked overall redundancy of the original form was reduced. (There was, understandably, strong pressure from the teachers to keep testing time to a minimum; but also to use whole-class periods for it. This necessitated some reduction and juggling of the test battery.) Sentences such as "I have pretty eyes" and "I am popular with boys" frequently caused disruptive behavior in the first year, and were therefore deleted. Because of the substantial change, comparisons between Year 1 and Year 2 are to be made using only those items common to both years, as a check on the analyses which use the entire instrument.

The instrument How This Class Makes Me Feel was added at the posttesting period in 1976. How This Class Makes Me Feel is a student self-report of attitude about the particular classroom in which he was observed. It was designed to assess the students' perception of the impact on self-esteem and social acceptance of the specific class in which they were tested and observed (in contrast to the full school experience reflected in P-H and SSD measures). The instrument consists of ten items, self-descriptive sentences, followed by five alternative boxes: NEVER, (BLANK), SOMETIMES, (BLANK), ALWAYS. The instructions read "The way I was treated in this class, this year, made me feel this way about myself." Phrases such as "It has made me feel good about my ability to do good work," "It has made me feel comfortable with my teacher," and "It has made me feel that the other people are happy to have me in the class," are typical items. Internal consistency and test-retest reliability coefficients have not yet been calculated for How This Class Makes Me Feel.

The School Sentiment Index administered in 1975-1976 consisted of 47 statements listed on two sheets attached to the two-page Student Self-Description. Seven items from the first year form had been judged to be inappropriate and were omitted. Some items were slightly reworded with the aim of clarifying the statements and simplifying wording so that students could understand the instrument without help. The total score for the SSI is used in the TLIS data analysis.

The Student Evaluation of Teaching instrument was revised after the first year of data collection. For Year 2, two items were omitted because both teachers and staff members found them inappropriate to the school situation. Some words were also changed to reduce reading difficulties. For the second year instrument the statements were placed first, each followed by three boxes. These boxes were labeled at the top of the page YES, SOMETIMES, and NO.

Changes in the Student Sentence Completion for Year 2 consisted of rewording certain items using a simpler vocabulary.

In the second year of data collection the Behavior Rating Scale was administered in fall and in spring. Pre- and posttest scores are available for BRS Self-Rating and BRS Other-Rating. The rating format in the second year instrument consisted of a single descriptive phrase with a five-point scale (NEVER - BLANK - SOMETIMES - BLANK - ALWAYS). Two items were dropped because of low item-total correlations and a new item was added to indicate response to peer pressure. For the posttesting in spring 1976 a thirteenth item was added: How Well Do You Know This Person, with two response alternatives - NOT AT ALL and VERY WELL .

As noted earlier, in both Austin and Kentucky in the first year of data collection, the Comprehensive Tests of Basic Skills were used as a measure of student achievement. CTBS was not used for all classes in the second year of data collection. The Metropolitan Achievement Tests (MAT) and the Gates - MacGinitie Reading Tests were introduced. Some classes were tested in Year 2 with CTBS. In Austin, two classes received CTBS Form S, Level 2 in math for both pretest and posttest because the participating teachers were teaching that subject. All other Austin classes were observed and tested in reading classes and used, therefore, the Gates-MacGinitie Reading Test. Ten of 25 classes in Kentucky were tested using CTBS Form Q, Level 3 in Reading Comprhension; the other 15 classes were tested with the MAT in language arts and reading because this test was being given by the state and it was not desirable to the schools to give two different tests.

The Gates-MacGinitie Reading Test, Survey D, Form 1, was developed by A.I. Gates and W.H. MacGinitie and is published by Teachers College Press of Columbia University of New York (Gates and MacGinitie, 1972). The reading achievement

test is standardized for fourth to sixth graders. The format is multiple choice, and it requires approximately an hour to administer. Two separate testing dates are suggested in the Gates-MacGinitie Administrators Manual, though the test was administered in one period for the TLIS. The instrument measures students' reading abilities on three scales: Speed and Accuracy, Vocabulary, and Reading Comprehension. For purposes of analyses with other TLIS data, the Gates-MacGinitie Reading Test is being scored on two scales: Cognitive, a combination of vocabulary and comprehension scales; and ACCRAT, a ratio of number correct to the number of items attempted. Internal consistency reliability of the Vocabulary and Comprehension scales are reported as .87 and .96 (split half-odd/even). The Gates-MacGinitie Technical Manual also cites validity coefficients of .78 and .80 for Vocabulary and Comprehension scales, respectively, by comparison of the Gates-MacGinitie Reading Test with four other standardized reading tests.

The Metropolitan Achievement Tests were developed by Durost, Bixler, Wrightstone, Prescott and Balow; they are published by Harcourt Brace Jovanovich, Inc. of New York. The MAT measures student mastery of study skills, parts of speech and punctuation, capitalization and usage. MAT Form F or G was administered to the Kentucky sample. Response format for each of the scales measured is multiple choice. MAT reports reliability coefficients for sixth graders as follows: total reading scale, .96 (Spearman-Brown and KR-20); word knowledge scale .92 and .93 (Spearman-Brown and KR-20); reading scale .93 (Spearman-Brown and KR-20); and language scale, .95 (Spearman-Brown and KR-20). MAT reports no validity coefficients (Prescott, 1973).

Some of the reading classes in Austin were provided with text-book-referenced Unit Tests as an additional measure of teaching and learning. A Unit Test is given after students complete a section of stories in a reading textbook. Some publishers provide Mastery Tests, as well, to cover material from the entire

book. The phrase Unit Tests is used in the TLIS to describe either a subtest or a mastery test. The following tests were used: Basic Reading Tests for Images and Galazies (Houghton Mifflin Co., 1974); the informal assessment test, the test for Exploring Afar, and the test for Discovering Treasure, from the Open Highways reading program (Scott, Foresman and Co., 1974); the Mastery Test to accompany Into New Worlds (Sipay, 1970); and Riders on the Earth (Holt, Rinehart and Winston, 1973).

Year Two, Teacher Instruments

Teachers did not complete new Biographical Forms, Views of Teaching, Adjective Self-Description, or Directed Imagination forms in the second year. Teachers who entered the project for the second year of data collection, however, were asked to complete those forms, with the exception of the Directed Imagination.

The Student Description instrument was altered only slightly for the second year. The items were not changed; the response format was. In the second year teachers were asked to describe the Special Study students as they were at the period of posttesting, using the 22 items of the SD. Teachers were also asked to rate each student on how much he or she had changed over the school year, on the same set of 22 items. Two scores are derived from the second year SD: SD-Status and SD-Change.

The School Preference Questionnaire and the Teacher Needs Survey were not administered in the second year of the project.

The Class Checklist, which had been administered with the Preference Questionnaire in the first year, was administered separately in the second year. The items of the Class Checklist were the same over both years. In the second year, however, teachers were asked to compare the observed class of the 1975-76 school year with the observed class of the 1974-1975 school year on the ten items. The Class Checklist is scored simply for a Similarity Index, the teacher rating of the degree to which the two groups of students were about the same.

In connection with the use of publisher-provided Unit Tests of reading skills, the Teacher Concerns Questionnaire was administered in spring of the second year of data collection. The Concerns Questionnaire was developed by Gene Hall, Archie George and William Rutherford following the work of Frances Fuller at the UT R&D Center, to measure teacher attitudes toward innovations. The questionnaire is in the form of a series of declarative sentences which the teacher is asked to rate, using a Likert-type scale of 1 - 7 according to how well the items describe him or her. Sentences include the following: I am concerned about students' attitudes toward this innovation; I am concerned about how the innovation affects students, etc. The instrument was adapted for use in the TLIS by attaching a cover sheet to the questionnaire to explain to teachers that phrases such as "This approach" and "The innovation" referred to their use of Unit Reading Tests. Since the questionnaire is used for a variety of innovations, no modification of the actual test was necessary. The instrument measures seven stages of concern: awareness, informational, personal, management, consequence, collaboration, and refocusing. Internal consistency reliability for the seven scales on the Concerns Questionnaire range from .64 to .83. Test-retest reliability ranges from .65 to .86 (Hall, George, and Rutherford, 1977).

In spring 1976 teachers who had participated in the Teaching-Learning Interaction Study were asked to respond to three items on the Teacher Reaction Form. The Reaction Form was utilized to provide feedback from teachers about the classroom observations that had been conducted during the year. The three items are: The observations were conducted at convenient times; the classroom situations observed were representative of the normal activities in my class; and the classroom observer did not detract from classroom activities nor the effectiveness of instructional activities. The response format is a five-point

Likert-scale that ranges from completely false to completely true; room for comment and an invitation to comment were provided.

Year 2, Classroom Observations and Teacher Interviews

Each of the observation and interview techniques was revised between the first and second years of data collection for the Teaching-Learning Interaction Study. The Classroom Observation Record was revised by adding scales to indicate the appropriateness of teacher's behavior and the accuracy of teacher's statements, bringing the total number of COR items to 30. The rating technique was not altered.

The Classroom Observation System was expanded considerably for the second year of data collection to allow more teacher and pupil behaviors to be coded and quantified for statistical purposes. In addition to student Time-on-Task, teacher On-Task Behavior was coded in the second year COS. Additionally, teacher and student Accuracy of Statement and Response, and Affect were coded. Student Reading Fluency and student Active/Passive behavior were also coded with the second year COS, as was teacher Appropriate Level.

The Classroom Observation Report Form (CORF) was changed for the second year. Item 30, Cognitive Level, was eliminated because first year data showed that Cognitive Level was characteristically "normal." A seventh space was added to Section I, Individual Students, to accommodate one additional Special Study pupil. At the right-hand side of the page, a column headed COR Rating was added. In this column the rater wrote the COR item numbers most closely corresponding to the behavior being described to the left of the number. In all, the revised CORF allowed the observer to provide more detail of teacher/pupil behavior and interaction than with the previous year's CORF version.

In the first year of the TLIS teachers were interviewed only once. In the second year of data collection teachers were interviewed five times by

interviewers. At least two interviewers conducted the total of five interviews. Three separate interview forms were used: the Initial Interview, the Interview Form and the Final Interview form. The initial interview was conducted during the week of the second observation session in the middle of the school year; biweekly interviews were conducted thereafter, in conjunction with classroom observations; and the final interview took place about a month before the end of the school year, just after the posttest period. In the second year of data collection interviewers also filled out the Interviewer Perception Form, following each interview. The Interviewer Perception Form allows the interviewer to evaluate the teacher in three areas related to the questions on the interview form: Diagnostic clarity and Precision, Prescriptive Skill, and Follow Through. To explain why a particular rating was given on the Interviewer Perception form, the interviewer illustrated each rating with an anecdote from the interview. Throughout the interviews, the interviewer recorded all the information the teacher gave, not just direct responses to the questions which were asked. The interviews are important in the body of TLIS data collected for the descriptive information which they yield concerning the Special Study students and the knowledge and skills of the individual teachers.

DESCRIPTION OF SAMPLE

In Austin the Teaching-Learning Interaction Study took place in four sixth grade centers. In the first year of the study 53 teachers participated; in the second year 42 teachers participated (one teacher was observed with two classes). Of those, 20 participated in Year 1, only, and 9 participated in the Year 2, only. In Kentucky, in four schools, 27 teachers participated in the first year and 25 in the second year of the study. There were 5 teachers who participated in the first year, only, in Kentucky and 3 who participated solely in Year 2. Demographic data about the teachers in the research sample include age, number of years of teaching experience, professional degree attained, ethnicity and socioeconomic status of the teacher's family of origin.

In the first year of the project, in the Austin schools, there were 1647 students participating. Of these, 69.5% (1145) were Euro -Americans; 14.9%(245) were Afro-Americans; and 15.6% (257) were Mexican-Americans. These ethnic groupings are referred to throughout the study, in the parlance they prefer in the Southwest, as Anglos, Blacks, and Chicanos. Of the sample 51.2% were males and 48.8% were females. Males and females are represented among the ethnic groupings in approximately those same percentages. The mean SES rating for Anglos was 14.3 (SD = 5.5); for Blacks, 18.9 (SD = 5.3); for Chicanos, 18.5 (SD = 5.2).

In the second year of data collection in Austin there were 1,513 students. Of those 75.2% (1132) were Anglos; 12.0% (180) were Blacks; and 12.9% (194) were Chicanos. Males represented 50.6% and females 49.4% of the population. Again, males and females are proportionally represented among the ethnic groupings. For Anglos in Year 2, mean SES was 13.3 (SD = 5.5); for Blacks mean SES was 19.6 (SD = 6.0); and for Chicanos, 18.1 (SD = 6.3).

SES is computed using a technique developed by Robert J. Havighurst for the Coping Styles and Achievement: A Cross-National Study of School Children,

on which he collaborated with Robert Peck. The SES score is based on the occupational category and educational attainment of the head of the household. The lower the score, the higher the SES level; social class estimates made with these scores are in relation to the breakdowns introduced by Lloyd Warner: SES score 5-12, upper middle class; 13-19, lower middle class; 20-25, upper lower class; 26-30, lower-lower class.

The median family income in Austin is $13,540; 18% of the population, however, live on incomes below the poverty level (Bureau of the Census, 1977). Thirteen thousand dollars is an upper middle class income in this city where the average age is about 25 and there is very little industry. Austin is the State Capital and the location of Texas' largest university.

According to the 1976 experimental census, Anglos comprise 70%, Blacks 12% and Chicanos 17% of the total population of the City of Austin, 308,952. Of this population, 4,668 are 11 year old children and 4574 are 12 years old. Children in the sixth grade are usually 11 or 12 years old. The numbers of 10, 11, 12, and 13 year old children in the city are quite similar; it is a fair assumption that proportional numbers of them are spread over grades 5, 6, and 7 and that comparison of our research sample's characteristics to any <u>one</u> of the age groups will yield representative information. Of both the 11 year old and the 12 year old population of the City of Austin, males comprise about 51% and females, 49%. Anglos make up approximately 60% of each group; Blacks are about 17% of the 11 year olds and 18% of the 12 year olds; Chicanos make up about 23% of the 11 year olds and 22% of the 12 year old children.

Our research sample in Austin represents one third or more of the sixth-grade-age children in the city. Anglos are slightly overrepresented for their age groups, although not for the city at large; Black children are slightly underrepresented, and Chicano children are more underrepresented, for their age groups, but not for their ethnic groups in the total city population.

The sixth grade centers were established in 1972 to end patterns of segregation that previously existed in the Austin school district. Students are assigned to classes in those sixth grade schools in a manner designed to insure proportionate representation of all groups of students. An equal number of girls and boys are assigned to classes. There is an attempt made to assign children to classes so that they are made up of approximately 63% Anglos, 19% Blacks, 17% Chicanos, and 1% Orientals (this latter group has been included in the practice only since 1976). The proportions follow DHEW guidelines. Additionally, the attempt is made to distribute students heterogeneously within and across classes, according to previous academic performance. Academic achievement is not a main criterion for assignment, but it is considered undesirable to group students into classes on the basis of achievement scores. The TLIS data are being analyzed for statistical evidence which speaks to the question of random or non-random assignment of students to classes.

In Kentucky, in the first year of the study, there were 783 students participating. Seven hundred eighty one of those were Anglos and 2 of them were Black. In the second year of the study children participated; were Anglos, and were Blacks. Mean SES for Kentucky students in Year 1 was (SD =); in Year 2, (SD =).

A stratified subgroup of students in each class, called the Special Study students, was selected for intensive observation in the classroom. Teacher concerns, actions and appraisal of these students were elicited during the teacher interviews. In the first year of the study there were 408 of these Special Study students in Austin. Approximately 75% of them were Anglos, 12.5% were Blacks, and 12.5% were Chicanos. In Kentucky in the first year there were 162 Special Study students. In the first year of the study, pupils were

selected to be Special Study students on the basis of their scores on the Behavior Rating Scale-Other Rating which was scored at that time for two factors, emotional adjustment and school coping behavior. One primary and one alternate student were chosen for the low, medium and high groups on each scale.

During the second year of the study in Austin, there were 290 Special Study students; % were Anglos; % Blacks, and % Chicanos. In Kentucky there were 165 Special Study students. In the second year of the study, pupils were nominated by their teachers to be in the Special Study subsample. Teachers used a nine-cell chart headed, Chart for Selecting "Special Study" Pupils, to nominate seven first-choice and seven second-choice students as representative of pupil types on two factors, Aptitude and Effort, using a range of high, medium and low on the two factors. Nine possible student types exist but only seven students were judged practical to observe, so that two of the "middle" types were eliminated. The seven pupil types were: High Aptitude/High Effort; High Aptitude/Low Effort; Average Aptitude/High Effort; Average Aptitude/Average Effort; Average Aptitude/Low Effort; Low Aptitude/ High Effort; and Low Aptitude/Low Effort. Teachers were asked to choose the two pupils who best represented each type within the class, based upon the teachers' experience with all six graders they have known.

During the second year of the study, teachers were asked to compare the two classes which were being tested and observed, on ten items which comprised the Class Checklist. The items assess both social and academic behavior. Austin teachers felt that their classes were similar about 26%; Kentucky teachers rated their classes' similarity at 42%. On the average then, teachers during the second year thought that the two classes were slightly less than 1/3 similar (30%) on the ten items of the Class Checklist.

TLIS PROGRESS REPORT

ANALYSIS OF DATA

THE ANALYTIC PLAN OF THE STUDY

OVERVIEW

The analytic procedures consist of these logically sequential steps (not necessarily in this order, operationally):

1) Psychometric analyses of individual instruments
   a. Item analyses, to eliminate unreliable or administratively troublesome items
   b. Factor analysis, to derive simple structure and best variable scores to be used in subsequent analyses
   c. Internal consistency reliability analysis (alpha)
   d. Test-retest reliability analysis
   e. Correcting for unreliability of measures
   f. Validity tests
2) Descriptive statistics on each pretest and each behavioral measure, testing for differences within each of these population subclasses:
   a) year sample; b) Site; c) school; d) class; e)age; f) sex
   g) SES level; h) ethnic group
3) Hypothesis-testing analyses: MANCHA - the nine-stage system

PSYCHOMETRIC ANALYSES

The specific procedures for steps (a) - (d), and the findings (to date), are described separately in the individual instrument manuals in Appendix A.

Step (e), the achievement of a "true score" correction for instrument reliability, was a particular concern from the outset of the study. Cronbach and Porter both urged that it be accomplished. When we tried the existing procedures, however, they turned out to create more false-positive results than an "uncorrected" analysis (Marston and Borich, 1977). Cohen investigated this further, for us, during 1976-77, and finally came to the same negative conclusion.

Cronbach confirmed this difficulty too in a recent letter of May, 1977. Consequently, all the time and analyses that have been devoted to trying to adapt or invent a usable reliability correction have proved unavailing. The state of the art had not reached a solution that was applicable to the multivariate model of this study by the time it began, and neither we nor our consultants have been able to invent a solution.

Just at this time, however, Noel Dunivant, whose interest in the project has continued to be lively and active, learned about Jöreskog's new method for making reliability corrections. In FY '78, we plan to investigate the applicability and validity of this method for our data. (See Appendix B, "Problems", below, for more details of this and other practical problems that have had to be met and overcome in the course of working through the project to this point).

Validity Tests

Seven kinds of validation tests are possible, using the total array of measures collected in this study.

1) Construct validity. This is tested by comparing similar or logically analogous measures of a given characteristic, derived from two or more instruments completed by the same pool of subjects. Examples include comparisons among the student self-report measures of attitudes (SSI, SET, SSC-A); comparison of the self-report measures of coping behavior (BRS-Self score, SSC-C); teacher attitudes (ASD-Attitude, VOT-A); etc.

2) Freedom from Ethnic or SES Bias (applies to ratings). Appropriate tests are performed in order to see whether a particular kind of rating has been subjectively biased by the ethnicity or SES level of the rater, the ratee, or both. The strongest test is to look for a significant interaction effect on the rating,of rater-status by ratee-status. If this is not present

(as in the case of the Austin, Year One, BRS-Other scores), it demonstrates that the rating is not affected by rater prejudice or bias. In this case, even if ratee groups differing on SES or Ethnicity get different mean scores on the rating scale, the pattern indicates that all judges agree that this difference is "real". They may, of course, all share a universal bias but there is no way to determine that from the data internal to the rating instrument. In the present study, it is possible to cross-check such a finding against independent behavioral measures, at a number of places. The issue is then tested by cross-validation across different measures, rather than across different judges using the same measure.

If bias is found in a rating, and particularly if the rating does not cross-validate substantially against other measures, then any further analyses that involve that rating must be interpreted with that bias--that questionable validity--explicitly taken into account.

3) Cross-Rater Validity. This is tested by comparing the evaluations of a pool of subjects by two or more sets of independent observers. Examples include comparisons of ratings of student behavior by peers (BRS-Other), self (BRS-Self), teachers (Student Description), and classroom observers (COR-Pupil scales); comparison of ratings of teachers by their students (SET, How This Class Makes You Feel) and by classroom observers (COR, Interview ratings); etc.

4) Behavioral Validity. This is tested by comparing either self-report scores or other peoples' ratings of a pool of subjects with behavioral measures on those subjects. These behavioral measures presumably represent the characteristics measured by the self-reports or ratings, or should be logically closely related to them. Examples include comparison of student self-

report of coping effectiveness (SSC-C, BRS-Self) with observer-coded classroom behavior records (COS: Time-on-task, Response Accuracy, etc.); comparison of teacher self-report of coping effectiveness (ASD: Efficiency, Anxiety, VOT-Coping) with observer coded classroom behavior records (COS: Time-on-Task, Accuracy of statements, Appropriate level of discourse, etc.); comparison of peer and teacher assessments of students' academic coping effectiveness with observer-coded behavior records (COS: student Time-on-Task, Response accuracy, Reading fluency); etc.

5) Consequent Validity. This is the converse of predictive validity. It tests the hypothesis that a given variable should logically be affected by certain other antecedent variables. For example, (assuming, for the moment, that the predictor measure is valid), student posttest achievement should logically be affected to a significant degree by student Time-on-Task behavior during the year. If it is not, this test of consequent validity would make questionable whether the achievement posttest was a valid measure of the kind of learning that the Time-on-Task process measure is logically presumed to affect. No one such test is taken as the basis for a final, definitive decision in this study; it is the total pattern of antecedent-consequent relationships that is found, for all antecedent variables and all consequent variables, that decides. If a preponderant amount of evidence indicates that a given variable, when treated as a consequent, fulfills the hypothesized pattern with most antecedent variables, then its validity might be taken as better established than the validity of one predictor measure to which it did not show the predicted consequent relationship. This would be particularly true if that predictor variable failed to affect some other "congruent" variables in the predicted manner.

This "consequent validity" concept has not often been invoked in a formal or systematic way, in past research. This probably is because of the

impossibility of avoiding circularity of reasoning when only one antecedent and one consequent variable are available, as is the case in univariate or bivariate studies. Perhaps only in multivariate studies such as the present one can this fallacy be avoided, by having several parallel, independent ways of testing the validity of any one variable when it is treated as a consequent.

6) Predictive Validity. This is tested by seeing whether a particular measure has a significant effect on change in another, independent characteristic. A post measure is taken at a later time than the predictor measure and there should be logical reasons to expect a cause-effect relationship. (This is desirable even though the purely observational, correlational nature of the present study permits no final proof of causality.) Examples include the study of the effect of a particular student characteristic (e.g. self-esteem) on regressed gain in achievement, in attitude, etc.; the effect of a teacher characteristic (e.g., COR-Kindly, understanding) on student achievement, attitude, etc.; and, particularly, in this research, study of the interacting effects of student and teacher characteristics on the student outcomes. Where teacher or student predictor variables are found to have significant effects on subsequent changes in students, this suggests they may be empirically useful. Particularly where such effects have been predicted in advance, confirmatory findings tend to imply that the predictor measures have an appreciable degree of validity. Where significant effects are found, but where they run contrary to the prediction, further research is necessary in order to ascertain whether the findings are chance results, whether the predictor measure taps something other than was originally supposed, or whether the theoretical formulation of the presumed cause-effect relationship was incorrect and needs to be modified in the light of the empirical evidence.

7) Replicative Validity. This is tested in two ways, one weak and one strong. Weak replication is found by comparing patterns of findings on logically analagous but independent sets of measures taken on the same population. An example would be a comparison of the effects of teacher attitudes (e.g., ASD-Attitude, or COR-Kindly, understanding) on, first, changes in student attitudes (e.g., Self-esteem) and, second, on changes in student coping skill (e.g., SSC-C). If the effects of teacher attitude prove to be similar on the two different outcomes, as predicted, this tends, somewhat distantly, to imply that the predictor measures did validly assess the characteristics they were presumed to assess.

Strong replicative validity is found when a similar pattern of antecedent-consequent relationships (or correlational relationships among simultaneous measures) is found in two or more independent samples. The present study is designed to replicate the analysis in two different parts of the U.S. and, in each place, to replicate the design with most of the same teachers but with different classes of students in two successive years. Thus, the data in each location are examined to see if a predictor-outcome relationship which is found in the first year is similarly present in the second year; and if the findings in Texas repeat the pattern found in Kentucky. Such analysis is now usually discussed as an aspect of generalizability theory. It does establish whether a finding appears likely to hold true "in general"-- or, more likely, for similar communities and similar school populations. It also tends, more distantly, to imply that the various measures employed do assess what they are logically supposed to assess. The fact that a whole web of interrelationships is stable across different samples seems to imply, though it cannot ever finally prove -- that the various parts of that web have been reasonably well understood, theoretically, and that its parts have been validly measured.

Where significant but different findings emerge in different samples, this may be due to chance, or to instability or invalidity of measurement. It can also, however, signal that there are genuine differences in the predictor-outcome relationships in different communities, or in different sub-sets of the population. While the other forms of validity tests can be addressed to this issue, further research is almost certain to be required to find the true explanation for the differences in the different samples. Limitations of funding, historically, have limited any such research to small, partial samples, mostly in the "tiny" range. The present study is no exception. It has more kinds of data, on larger, more diversified samples than most previous studies; but it is no more than a pilot-size venture when a definitive understanding of learning in the total American school population is the end in view.

DESCRIPTIVE STATISTICS

Appropriate descriptive statistics are generated for the eight levels of population sub-groups listed above. Some of these are a by-product of analyses performed in the nine-stage model; some are independently calculated. In either case, tests for significant sub-group differences are performed. In addition to means, standard deviations and other common descriptors for each pretest and each behavioral process measure (for both student and teacher variables), certain additional information is also generated and examined: for example, the pretest-posttest correlations on the student measures. The fact that Black children have a pre-post correlation of only .34 on CTBS Achievement (Austin, Year One), whereas Anglos and Chicanos have correlations of .75 and .72 , respectively; and the further fact that Black students have much higher pre-post correlations on other instruments, at or near the level for the other two ethnic groups--these facts raise important questions about the ethnically differential reaction to the standardized achievement test

which urgently require further inquiry. Clearly, it is not a question of differential literacy generating more randomness of pre-post answering among the Blacks, or their pre-post correlation on other instruments would be similarly low. Something peculiar to the achievement test creates this effect. What is it?

In the course of the analysis of the data, other kinds of statistical descriptors may also turn out to show significant group differences. They will be added, when identified, as further data to describe each group's special characteristics.

HYPOTHESIS TESTING

One major issue in the TLI Study is to study interactions between teacher and pupil characteristics that may produce change in students over the course of the school year. The "main effect" relationships of both student and teacher characteristics to student change are equally important questions, too. In order to study change, it is necessary to describe the state of the students at the beginning of the year and what factors might have affected their pretest scores. The data collected for the TLI study have been described. Figure 5 shows the instrument-variables that were subjected to the analysis described here. These measures are all continuous variables, except for sex and ethnicity. While the total Austin, Year One sample, for example, included 1,657 students in 53 classes, the number used in any particular analysis shrank somewhat, due to the student or the class having missing data on a particular variable. The average number of students was generally 1,100 to 1,200, in 51 to 53 classes, for the analysis of any one variable, using this sample.

In order to deal effectively with this rather large body of data, and to answer some fairly complex questions, a general linear model approach was used, with an emphasis on hypothesis testing. The general approach used was to set up mini analyses each of which had one criterion.

Generally, the students' entering characteristics, as measured by the pretests, were used as predictors, as were the teachers' personality and classroom behavior measures. Using separate analyses, both the student pretest and the student posttest were used as a criterion. The reason for using the pretests as criteria was to find patterns of relationships among the students' personality traits present upon entering the classroom situation. The posttest measures were used as criteria to determine what pupil or teacher characteristics, of their combinations, would be most beneficial to student change over the

Variables Used in the Eight Stage Analysis

Up to September 1977

Pupil		Teacher
Pre	Post	
SET	SET	COR KU
SSC A	SSC A	COR SO
SSC C	SSC C	COR SI
P-H	P-H	Classroom means on pupil variables
CTBS	CTBS	
SSI	SSI	
BRS		
Ethnicity		
Sex		
SES		

Figure 5.

Footnote: All other teacher characteristics (ASD, VOT, demographic), teacher behavior (COS), and pupil behavior (COS, COR-P, SD) measures will ultimately be processed through the nine stage system.

year on a given outcome. When a posttest measure was used as a criterion, the pretest measure was always used as a covariate, in an attempt to remove the influence of the initial pupil level from the measure of posttest level; thus, the pretests were assessed for their contribution to regressed change.

Null hypotheses were tested. For example, one null hypothesis would be that SSC-A pretest level does not predict SET change. This kind of hypothesis test would minimally require the following two models:

1) $Y = a + bX + cS + E^{(1)}$ and

2) $Y = a + bX + E^{(2)}$,

where Y is the SET posttest score; X is the SET pretest score; and S is the SSC-A pretest score. If model 1 predicts significantly more of the variance than model 2, the null hypothesis is rejected and SSC-A is said to have a main effect on SET change.

Once linear models, and especially analysis of covariance, had been chosen as the method of analysis, several decisions had to be made. First of all, when continuous variables are used, some decision must be made as to the type of relationships that are allowed to exist (i.e. linear, quadratic, cubic). Optimally, one would like to represent the data as accurately as possible, while keeping the scope of the analysis within reasonable limits. In many studies, only linear relationships are studied but, in this study, the investigators felt that quadratic relationships were worth pursuing since it is not uncommon to find curvilinear relationships (Soar, 1975). In general, quadratic components were tested in early phases of the analysis. If they proved insignificant, they were dropped from subsequent analyses.

Another problem that had to be faced was the problem of missing data. Most of the 1,657 students had missing data on at least one variable. The general approach to this problem used here, was to include only subjects having

data for all the variables in a given "mini" analysis. In the early stages, however, students with missing data were included and a prediction was added which divided the subjects into two groups, those with data on an instrument and those without data on that instrument. This prediction could then be tested for significance. These tests were used to determine the appropriateness of later analyses which ignored subjects with missing data.

Finally, the analysis of covariance, which is used here to study regressed change, requires that the pre-post regression be the same for all values of the other predictors in the model. This is commonly referred to as the assumption of homogeneity of slopes and in many studies is simply assumed to be true. In these analyses, this assumption is tested. Adding these tests to the plan increases the complexity of the analysis, but this is more than compensated for by the increased confidence one can have in the results.

Multivariate Analysis of Change (MANCHA)
The Nine-Stage Analytic System

This section presents a synopsis of the 9-stage analysis system, and how the information from each stage uses and augments the information obtained from earlier stages. Conceptually, the primary variables fall into three categories: (1) teacher variables, (2) pupil variables that describe the pupils' entering condition, including demographic status and pretest measures; and (3) posttest measures, describing the pupils' state at the end of the school year.

Stages I, III, and VI are designed primarily to give information about the pretests. This information is used, first of all, to determine if different groups of students begin the school year at the same level on a variable. For example, if a significant ethnic difference on a pretest variable appeared, the standard analysis of covariance model, which is used

later, would be called into question, since one of the assumptions is a relatively equivalent pretest position of all student groups. This particularly affects the interpretation of ethnic differences in regressed gain.

A second reason for analyzing the pretests is to determine their interrelatedness. For example, if attitude and self-esteem pretests were highly related, one might look particularly to see if the regressed gains (to posttest) on these variables were related, also.

Finally, the pretests are analyzed with respect to the teacher and pupil characteristics in Stages VI and VIII, in order to determine whether there was, in fact, random assignment of students to particular classrooms.

The sequence of analysis of regressed change (adjusting posttest for pretest level) goes from analysis of main effects in Stages II and IV, to analysis of teacher-by-student interaction effects in Stage V, and thence to a further screening and refining of these preliminary findings in Stages VII, VIII, and IX. If one pupil pretest significantly predicts gain on another pupil variable, it is tested in Stage VIII and in Stage IX, against all other predictors (class, teacher, and student),to determine whether it has a unique predictive power. Similarly, a significant teacher predictor of gain in Stage IV is tested in Stage VIII, using the class as a unit of analysis, to determine whether that teacher predictor uniquely determines a significant proportion of the pupil change variance.

If a significant teacher-by-pupil pretest interaction predictor is found in Stage V, it is screened in Stage VII, using the class as a unit of analysis and taking the class-by-student variance as the error term. This usually provides a more conservative test of the hypothesis that there is a teacher-by-pupil interactive effect on student change.

Stage VIII tests for the unique main effect of a specified teacher or pupil characteristic, in the presence of all the other predictors. This gives important supplementary information on the amount of predictive redundancy in the student and teacher measures. Finally, in Stage IX, all of the significant predictors of change on a student variable are tested against each other to determine possible redundancy among the significant predictors. Thus, the nine-stage analytic system is designed to elicit the maximum amount of information concerning the precise effect of various factors on pupil change. Using more complex models in the later stages, it systematically screens those effects found in the simpler, earlier stages, to be sure they are valid findings. At the completion of the nine stages of analysis, the following results are examined: 1) the relationships among the pupil characteristics at the start of the year; 2) the student and teacher characteristics which individually predict, and therefore possibly affect, pupil gain or change on each outcome measure; 3) teacher by pupil interactions which predict, and therefore possibly affect, pupil gain on each outcome measure; 4) the unique predictive effects of the various measures, and their combinations, on student gain; and 5) the predictive effects of teacher, and teacher-by-pupil predictors when the classroom, rather than the individual student, is the unit of analysis.

Short Descriptions of Stages I - IX

1) Stage I is designed to determine whether or not students of different ethnic background or socioeconomic status (SES) differ on the pupil pretest measures; includes quadratic terms to test curvilinearity.

2) Stage II tests for possible main effects and interaction effects of ethnicity and SES on the pupil posttest (outcome) measures, with the pretest used as a covariate.

3) Stage III tests the correlations among the student pretest scores for significance.

4) In Stage IV, the linear and quadratic components of the student and teacher pretests are used to predict each student posttest, with the pretest used as a covariate. The objective is to find whether a teacher or student measure is, by itself, predictive of outcome.

5) Stage V tests the teacher-by-pupil interaction effects for their ability to significantly predict the pupil posttest scores on each outcome, with the pretest used as a covariate. A practical implication of a significant interaction effect is to suggest which students might best be assigned to which teachers, for the purpose of optimizing student change on a particular outcome measure.

6) Stage VI identifies any teacher-by-pupil pretest interaction effects in predicting a student pretest. Basically, this stage is used to uncover any systematic bias in classroom assignment. Stage VI is performed only on those variables which showed a significant interaction effect on the posttest in Stage V.

7) In Stage VII, the significant interaction findings from Stage V are further analyzed, using the classroom rather than the student as the unit of analysis. This stage is designed to check the possibility that students in a class share a common influence, and therefore that their scores are not truly independent observations (an aspect of the "nested variance" problem).

8) Stage VIII is designed to find if any of the instruments have a unique effect in predicting change. The unique predictive power of each student and teacher pretest measure is assessed by finding whether it predicts the student posttest criterion in the presence of the other predictors. Further, the teacher effects are tested using the class means as the unit of analysis.

Finally, the teacher effect on the pretest is measured.

9) Stage IX is yet to be done as of September 1977. It will determine whether the interaction effects found in Stages V and VII are uniquely predictive in the presence of all the other variables which are also found to predict a particular student posttest criterion.

SOME PRELIMINARY FINDINGS

Figure 6 summarizes, necessarily in a very sketchy way, the weight of the evidence to this point, as it bears on the twenty general propositions of the study. The specific evidence from the analyses so far completed on the Austin Year 1 and Year 2 samples, is presented in the discussion that follows. So far, each proposition (except #3 and #19) has been partially or wholly confirmed to a significant degree, although not for every variable and not always in both samples. The empty spaces for Year 2 are where analyses have not yet been run.

These propositons are quite general, of course. This was intentional. By stating them at this level, it is possible to test whether there is a general,lawful connectedness in the effects of teacher and student characteristics and behavior, taken singly and together. Considering the weakness of the measures, and the amount of "noise" from unidentified or unmeasured sources in the schoolrooms under study, even the present amount of convergence suggests that true relatedness does exist between teacher personality and teacher behavior; between child personality and aptitude, and child behavior; and between the interplay of teacher and child, and the changes in student knowledge, coping skills and attitudes that take place over the school year.

Just as important, of course, are the substantive details of the particular relationships that emerge; these are what speak to the practical decisions that teachers and other educators must make. If a pattern replicates again and again, it might safely suggest an instructional prescription, whether for all teachers and all children, or differently for different kinds of children or teachers.

If, however, a substantial pattern appears in any one sample but the specific patterns are different in different samples, this <u>may</u> be evidence

A SUMMARY OF PRELIMINARY FINDINGS OF THE TWENTY GENERAL PROPOSITIONS OF THE STUDY

Proposition	Analysis Stage	FINDING Year 1		Year 2	
1. Student entry characteristics will intercorrelate positively	III	Yes		Yes	
2. Entry level on one will affect change on the others	IV	Yes		Yes	
3. Positive teacher characteristics will positively affect teaching behavior	Like III	Preliminary: No		Preliminary: No	
4. Positive teacher characteristics will positively affect student behavior	Like III	TOT: Yes			
5. Positive teacher characteristics will positively affect student outcomes	IV - IX	Yes			
6. Positive teaching behavior will positively affect student behavior	IV - IX	TOT: Yes			
7. Positive teaching behavior will positively affect student outcomes	IV - IX	Yes			
8. Positive student behavior will positively affect student outcomes	IV	TOT: Yes			
9. Teacher/pupil interaction effects will influence student outcomes	V,VII,IX	Yes		V - Yes	
10. Each outcome will be produced by a unique set of factors	VIII, IX	VIII-Yes			
11. SES will correlate positively with student entry measures	I	Yes: Achievement, BRS-OR, Self-Esteem	No: SET, SSC-C, SSC-A	Yes: Achievement, BRS-OR, SR, Self-Esteem	No: SET, SSC-C, SSC-A, SSI
12. SES will positively affect student outcomes	II	Yes: Achievement, SSC-C, Self-Esteem	No: SET, SSI, SSC-A	Yes: Achievement	No: SET, SSI, SSC-C, A, Self-Estm
13. Ethnic differences will appear in student entry measures	I	Yes: Achievement, BRS-OR, Self-Esteem	No: SET, SSI, SSC-A, C	Yes: Achievement, BRS-OR, SR	No: SET, SSI, SSC-A, C, Self-Estm
14. Ethnic differences will affect student outcomes	II	Yes: Achievement, SSC-C, A, Self-Esteem	No: SET, SSI	Yes: Achievement, SSC-C	No: SET, SSI, SSC-A, Self-Estm
15. Many of these relationships (1-14) will be curvilinear	ALL	YES		NO	
16. Each student instrument will cross-validate (up to 7 ways)		Preliminary: Yes		Preliminary: Yes	
17. Teaching behavior will be validly measurable	Like III	Yes		Preliminary: Yes	
18. Individual teaching behavior will prove relatively stable	Special	Yes		Preliminary: Yes	
19. Correction for instrument reliability will improve accuracy	Not yet: JÖreskog method will be tried				
20. Findings will replicate across samples (although some may not, because of real differences in samples, and their reactions, beyond the level of chance fluctuations)	No: Many specific interaction effects, e.g., are unique to each sample. Yes: Many main effects of student characteristics replicate.				

Figure 6

that "prescriptions" cannot simply be tailored in advance, without knowing a lot more about a particular class, or a particular child. The final conclusion of such scientific inquiry might be that the art of teaching has to remain a highly skilled art, with judgment about how to teach a given child, or class, made at the time and point of contact, not a long time before. But that is anticipating... The results of the study, thus far, are as follows. They are incomplete and preliminary. The final results will be available in about a year.

Proposition 1

Student entering levels on achievement, coping skill, attitudes toward school and life, and self-esteem will all correlate positively with one another.

Year One: Of twenty-one correlations among seven variables, all but five accounted for more than 1% of variance. The attitude and self-reported coping variables (SSI, SSC-A, SET, SSC-C) clustered together. Peer-rated coping skill and achievement correlated substantially in a second cluster, linked to the first through school attitude (SSI) and self-esteem. Self esteem was positively related to all of the other measures, about equally. Significantly, evaluation of the teacher (SET) was not affected by the student's achievement or coping reputation. High achieving students did not rate their teacher any differently than did low achieving students.

Year Two: Of the twenty-one correlations, (aside from the replicated lack of correlation of SET with achievement and coping reputation), only one contributed less than 1% variance. Most correlations were substantially larger than in Year One. (See Table 1).

Preliminary interpretation: So far in the study, the proposition has been verified.

TABLE 1

CORRELATIONS AMONG THE PRETEST MEASURES

Austin, Year One and Two

	SET PRE		SSC A PRE		SSC C PRE		P-H PRE		BRS		CTBS PRE	
Year	1	2	1	2	1	2	1	2	1	2	1	2
SET PRE												
SSC A PRE	.22	.33										
SSC C PRE	.19	.28	.53	.52								
P-H PRE	.18	.21	.38	.23	.36	.36						
BRS	.05	.01	.16	.13	.13	.25	.26	.23				
CTBS PRE	-.01	.02	.07	.06	.07	.13	.26	.28	.42	.36		
SSI PRE	.45	.50	.54	.56	.45	.52	.39	.49	.21	.21	.09	.18

Proposition 2

Student achievement level, coping skill, attitudes and self-esteem will have reciprocal, positive effects on change in one another, over the school year.

Year One: Stearns (1971) summed up much evidence, "For disadvantaged children, school achievement has been observed to decline, relative to norms of performance, as number of years in school increase." Indeed, even in those studies where enriched preschool experiences lessened their disadvantage, this benefit disappeared in most cases after they were in elementary school for two or three years. Weinberg's review (1977) essentially confirmed the progressive widening of the gap in reading and other skills, between White majority children and Black, Spanish-speaking, and Indian children, over the twelve grades. The national "Assessment of Reading Activities" (1975) added further, similar evidence.

Stearns also found that "effects...on children's social and emotional development are not well researched." Although her statement concerns preschool effects, it describes equally well the state of research on the later years of school. In the only study of this kind we could locate, Stanwyck and Felker (1973) found that across grades 3-6 (cross-sectionally) "pupils with low self-concept gradually assume less responsibility for school success, while self-respecting assume more." The authors recommended further longitudinal research.

The Derivation of Table 2A and 2B

The following equation (1) represents the logical model utilized in Stage IV (pupil effects) analysis. The elements of the equation correspond to the effects tested in this stage. Tables 2A and 2B display only those

effects which met the following criteria, although all the effects were tested: (a) $p \leq .01$ and percent of variance explained $\leq$ 1.0 ; and (b) $.05 \geq p \leq .01$.

$$Y = X + P^L + P^Q + X^L * P^L + X^L * P^Q + X^Q * P^L + X^Q * P^Q \quad (1)$$

where: Y - criterion variable - SET Post, SSC-A post, SSC-C post, CTBS or GMG3 post, SSI post, and P-H or SSD, post.

X - pretest of criterion available

P^L - linear component of pupil predictor - SET pretest, SSC-A pretest, SSC-C pretest, CTBS or GMG 3 pretest, SSI pretest, P-H or SSD pretest, BRS-OR pretest, and BRS-SR pretest

P^Q - quadratic component of pupil predictor

$X^L * P^L$ - criterion pretest-linear by pupil predictor-linear interaction

$X^L * P^Q$ - criterion pretest-linear by pupil predictor-quadratic interaction

$X^Q * P^L$ and X^Q by P^Q - criterion pretest-quadratic by pupil predictor-linear; and criterion pretest-quadratic by pupil predictor-quadratic

There are six outcome measures, with six student predictors for each one in Year 1 and seven student predictors for each one in Year 2, making 36 and 42 main effect tests for Year 1 and Year 2 respectively. Of the 36 main effects tests in Year One, 31 were significant. As for Year 2, 17 out of 42 main effect tests were significant. These significant effects are not necessarily independent of each other because there are interrelationships among the outcome measures. However, the results, taken as a whole, are far beyond chance.

Year One:

The major results can be summed up as follows: (1) Students who started out high on any one attitude measure (SET, SSI, SSC-A) tended to gain more over the year in all other attitudes, and in coping skill (SSC-C), than did students with less favorable initial attitudes. The reverse was also true: high initial coping skill produced more favorable attitude changes than low coping skill. This was a circularly reinforcing 'closed loop'. (2) Students who started out high in achievement gained more in self-esteem, while low initial achievers actually lost in self-esteem. Conversely, students high on initial self-esteem made greater achievement gains than those starting lower. This was another circularly reinforcing 'closed loop'. (3) Students who scored high on initial achievement gained more in attitudes (except for SET, which is an evaluation of a particular teacher) and more in self-rated coping skill, than those initially low on achievement. Similarly, students initially high on self esteem gained more on all other outcome measures than those starting with lower self-esteem;" those high on SSC-A and SSC-C also gained in self-esteem: another 'closed loop".

The exceptions to this pattern of reciprocal reinforcements were that initial attitude scores did not affect achievement gains, and neither SSI nor SET initial levels affected change in self-esteem.

A pattern very similar to that described by Stearns is evident in these results. The gap among students in achievement appeared to be widening, in conjunction with their self-esteem; and, further, the gap among students in attitudes and coping skills also appeared to be widening. Those students who were psychologically and intellectually 'rich' to begin with, made greater gains than those who started out 'poorer'. Gains on academic achievement were

affected by self-esteem, but not by attitudes toward school, teacher or life in general.

"Rich-poor" is not just a catch-phrase, however. Minority and low SES students tended to score lower on initial achievement and self-esteem measures, even though they suffered no such disadvantage on the attitude and coping skill measures. This finding suggests that many minority and low SES students might be trapped, from the beginning, in a downward spiral, where decreasing achievement and decreasing self-esteem have negative effects on each other. If this pattern of results found in the sixth grade were true in earlier grades and if it continued in later grades, so that the cycle operated over a number of years of schooling, it would be easy to see why "the rich get richer and the poor get poorer," in many academic and personal respects, by the twelfth grade. The Year Two sample supplies the first cross-check on this finding.

Year Two:

The "closed loop" between attitudes and coping skill, found in Year One, was mainly but not entirely replicated in Year Two. Students who started out high in any of the three attitude measures (SET, SSC-A, SSI) tended to gain more in self assessed coping skill (SSC-C) than students with less positive initial attitudes. To close part of this loop, those initially high on SET, SSI, or SSC-C gained more over the year in attitude toward life (SSC-A); and students higher on SSI gained more in attitude toward teacher (SET).

A reversal of the "loop" effect appeared, however, where students starting low on SET, SSC-A, and SSC-C gained more in attitude toward school (SSI) than those starting higher. Thus, the picture in the Year Two sample is more complex than in Year One, and contradictory in one part.

Where achievement and self-esteem formed a mutually reinforcing "closed loop" in Year One, this did not reappear in Year Two. Initial achievement level predicted change in self-esteem but only in a very weak and complex way (including all main and interaction terms). Initial self-esteem did not affect achievement gain.

As for the third "loop", where SSC-A and SSC-C and self-esteem were largely reciprocal in their effects in Year One, in the second sample, this circularity disappeared. Self-esteem did relate positively to gain on SSC-A and SSC-C but the converse was not true. As in Year One, change in self-esteem was not affected by initial SET or SSI score; but now the positive effect of self-esteem on SET and SSI changes disappeared.

Preliminary interpretation: Initial status on one characteristic did affect change on a number of other characteristics, in both years (see Table 2A & B). This confirms the original proposition in the most general sense. Exactly the same effects appeared in both years for most predictors, on the SSC-A and SSC-C; many of the findings replicated on the SSI. Just two findings were the same for the achievement measures however, and there was practically no similarity of findings on the SET or self-esteem measures. There was a beyond-chance number of significant effects in both years on the achievement, self-esteem, and SET measures but the relationships were different in the two different samples.

It is important, therefore, to note the major differences in the specific relationships in the two samples. In particular, the disturbing uniformity of the "poor getting poorer" syndrome in Year One did not reappear in Year Two. Many possible reasons for such non-replication were discussed earlier. Nonetheless, the Year Two findings indicate that the outlook is not inexorably foredestined, and is not always as predictable from any one or two entry measures as was true in the Year One sample.

It is just such discrepancies, with practically important implications, that call for further research into the precise reasons how and why things can work (or appear to work) so differently in two different samples.

It will be interesting to see which, if either, of these patterns may replicate in one or both of the Kentucky samples. It will also be of interest to see what patterns emerge when this kind of analysis is ran separately for each of the three ethnic samples in Austin, comparing them with each other and with the Kentucky (Anglo) samples.

Interaction effects: Two types of interaction patterns are visible in the data summarized in Tables 2 and 6-11, which must be taken into account before interpreting any of the significant main effects reported in Table 2. The first is when the effect of another variable depends on students' level of score on the criterion pretest. For example, in Year One, the positive linear effect of initial peer coping reputation (BRS-O) on achievement gain was moderated by pretest achievement level. The data in Appendix B-4, Table 4,, show that BRS-O was positively correlated with achievement gain for students who were high or average on the achievement test in September; but initial BRS-O did not make any visible difference in achievement gain within the group of students who were in the lowest third on achievement in September. Thus, the "finding", relationship of BRS-O to achievement gain, was not true for initially low achievers.

In Tables 6-11, there are a number of places, particularly in Year One, where the effect of a pupil predictor variable on another student outcome is modified by interaction with a teacher measure. For example, in Table 2, it can be seen that change in attitude toward school (SSI) was positively related to pupils' initial self-esteem (P-H). Table 8A, however, shows that change in SSI was affected by an interacting combination of pupil self-esteem and

teacher kindliness (P-H * COR KU). Table 15a in Appendix B-5 shows that pupils average in self-esteem were more strongly affected by Teacher Kindliness. Teachers low on KU had a relatively stronger negative effect on their attitude, over the year, as compared with high KU teachers, than was true of pupils who were low in initial self-esteem. The overall pattern did indicate that low KU teachers had a more depressing effect than teachers higher on KU, for all pupils, but with the qualification just cited.

Table 2A

Summary of the Stage IV (Pupil Effects) in the Austin Sample, Year 1 and Year 2

(The findings listed below met the double criterion of $p \leq 0.01$ and percent of variance explained $\geq 1.0\%$ in either or both years).

CRITERION: SET REGRESSED CHANGE

Year 1				Year 2			
Effect[1]	P Level	% of variance	direction of effect[2]	Effect[1]	P Level	% of variance	direction of effect[2]
SSC-A[L]	<.001	2.2	+				
SSC-C[L]	<.001	1.2	+				
P-H[L]	<.001	1.4	+				
BRS-OR[L]	<.001	1.9	+				
SSI[L]	<.001	3.7	+	SSI[L]	<.001	1.43	+
CRITERION: SSC-A REGRESSED CHANGE							
SET[L]	<.001	1.7	+	SET[L]	<.01	0.93	+
SSC-C[L]	<.001	1.3	+	SSC-C[L]	<.001	2.82	+
P-H[L]	<.001	3.1	+	SSD[L]	<.001	3.11	+
BRS-OR[L]	<.001	2.7	+	BRS-OR[L]	<.001	1.48	+
CTBS[L]	<.001	1.6	+				
SSI[L]	<.001	3.3	+	SSI[L]	<.001	4.95	+
not reported in year one				BRS-SR	<.001	1.94	+
CRITERION: SSC-C REGRESSED CHANGE							
SSC-C pre-missing pretest vector	<.01	0.5		not run in year two			
SET[L]	<.001	1.1	+	SET[L]	<.001	1.38	+
SSC-A[L]	<.001	1.9	+	SSC-A[L]	<.001	2.23	+
P-H[L]	<.001	5.0	+	SSD[L]	<.001	1.50	+
BRS-OR[L]	<.001	4.8	+	In Table B (p=.05-.01, V=0.61%)			
CTBS[L]	<.001	2.4	+				
SSI[L]	<.001	4.0	+	SSI[L]	<.001	5.21	+
not reported in year one				BRS-SR[L]	<.001	3.25	+
CRITERION: ACHIEVEMENT REGRESSED CHANGE							
SES[L]	<.001	1.8	+	SES[L]	<.001	2.51	+
P-H	<.001	1.2	+				
BRS-OR[L]	<.001	2.7	+	BRS-OR[L]	<.001	2.52	+
X by BRS-OR	<.001	1.2					
CRITERION: SSI REGRESSED CHANGE							
SET[L]	<.001	4.3	+	SET[L]	<.01	1.17	-
SSC-A[L]	<.001	1.7	+	SSC-A[L]	<.001	1.97	-
SSC-C[L]	<.001	1.1	+	SSC-C[L]	<.01	1.16	-
P-H[L]	<.001	2.3	+				
BRS-OR[L]	<.001	2.3	+				
CTBS[L]	<.001	1.9	+				

Table 2A

CRITERION: SELF ESTEEM REGRESSED CHANGE

Year 1				Year 2			
Effect[1]	P Level	% of variance	direction of effect[2]	Effect[1]	P Level	% of variance	direction of effect
SSC-A^{L}	<.001	1.0	+				
SSC-C^{L}	<.001	1.0	+				
BRS-OR^{L}	<.001	2.8	+				
CTBS^{L}	<.001	3.5	+				

1
L = linear
Q = quadratic
X = pretest of criterion variable
X^{M}= missing pretest data
P-H^{M}= missing predictor data

2
\+ = positive direction
\- = negative direction

Table 2B

Summary of Stage III, (Pupil Effects) in the Austin Sample, Year 1 and Year 2

(The findings listed below met the $.05 \geq p > .01$)

CRITERION: SET REGRESSED CHANGE

Year 1				Year 2			
Effect[1]	P Level	% of variance	direction of effect[2]	Effect[1]	P Level	% of variance	direction of effect
X by SSC A	< .05	0.23					
P-H[Q]	< .01	0.5					
X by P-H	< .05	0.32					
CTBS	< .01	0.6	+	CTBS	<.01	0.86	
X by CTBS	< .05	0.26					
X by SSI	< .05	0.29		X by SSI	<.05	0.51	
not reported-year 1				BRS-SR[L]	<.05	0.56	
not reported-year 1				BRS-SR[Q]	<.05	0.47	

CRITERION: SSC-A REGRESSED CHANGE

				X by GMG 3 and X by GMG[Q]	<.01	0.84	
				X by GMG 3	<.01	0.73	
SSI[Q]	< .01	0.36					
not reported-year 1				X by BRS-SR	<.05	0.44	

CRITERION: SSC-C REGRESSED CHANGE

				SET[Q]	<.05	0.47	
P-H[Q]	< .001	0.8					
X by P-H	< .05	0.39					
In Table A				BRS OR[L]	<.05	0.61	+
X by BRS-OR	< .05	0.26					
X by CTBS	< .05	0.36		X by GMG 3	<.01	0.60	
				X by GMG 3 and X by GMG 3[Q]	<.05	0.62	

CRITERION: CTBS OR GMG 3 POST

X by SSC-A	< .05	0.21					
X by SSC-C	< .01	0.31					
X by P-H	< .01	0.26					
BRS-OR[Q]	< .01	0.6					
X by BRS-OR and X by BRS-OR[Q]	< .001	0.93					
SSI[L]	< .05	0.45	+				
X by SSI	< .01	0.27					

Table 2B

Summary of Stage III, (Pupil Effects) in the Austin Sample, Year 1 and Year 2

CRITERION: SSI

Year 1				Year 2			
Effect[1]	P Level	% of variance	direction of effect[2]	Effect[1]	P Level	% of variance	direction of effect[2]
X by SET	<.001	0.6					
				X by SSD and X by SSD^Q	<.05	0.55	
X by P-H	<.05	0.22		X by SSD	<.05	0.38	
X by BRS-OR	<.05	0.29					
not yet run in year 1				$BRS\ SR^L$	<.01	0.81	-
not yet run in year 1				X by BRS-SR	<.05	0.31	

CRITERION: P-H or SSD REGRESSED CHANGE

SET^Q	<.01	0.8					
X by SET and X by SET^Q	<.001	0.88					
X by SET	<.001	0.55					
				SET^L	<.05	0.56	
				X by SSC-C and X by $SSC\text{-}C^Q$	<.05	0.57	
BRS OR (missing data vector)	<.01	0.6		not yet run in year 2			
CTBS (missing data vector)	<.05	0.38		not yet run in year 2			
SSI^L	<.05	0.35	+	$GMG\ 3^L$	<.05	0.42	+
				X by SSI and X by SSI^Q	<.01	0.78	
X by SSI	<.05	0.26		X by SSI	<.05	0.76	

1
L = linear
Q = quadratic
X = pretest of criterion variable
X^M = missing pretest data
$P\text{-}H^M$ = missing predictor data

2
+ = positive direction
- = negative direction

Proposition 3

Positive teacher characteristics will positively influence teaching behavior.

Generally, this postulates that the traits which are (probably weakly) estimated by such self-report instruments as the ASD and VOT, will shape the kinds of teaching behavior that are estimated by the COR and COS measures generated by classroom observers. The analyses completed thus far that bears on this proposition is Pearson correlation of the ASD and VOT scores with the COR scores (all class means), with the results shown in Table 3. Except for negative correlations of self-rated attractiveness with COR Stimulating-inventive, there were no relationships large enough to be significant. The proposition, so far, is disconfirmed. Correlation of ASD and VOT with the COS measures awaits analysis in the near future. The lack of simple correlational relationship of the teacher self-report instruments and the classroom observations may be somewhat misleading, however, since both the self-report measures and the COR measures showed a sizable number of significant relationships to student behavior and to student gains. These relationships were modest in size but substantial in number. (See Propositions 4, 5, 6 and 7 below) The observers,or these rating factors,may have "missed" something the students picked up and reacted to. Another possibility is that the self-report scales and the ratings measure two relatively unrelated sets of teacher properties, each of which has an influence on students.

The one other significant finding in Table 3 is the negative correlation of COR-Kindly- understanding with teacher Age in Year 1. (Year 2 has not yet been analyzed on this point.) The Age effect is strongest and probably accounts for most of the other two correlations. The observers found that

Kindly-understanding behavior tended to fall off with increasing age in a good many cases. This may be a pertinent place to report that teacher Age correlated .25 with ASD-Efficiency and .28 with ASD-Anxiety. Years of Experience correlated .37 with ASD-Efficiency and .30 with ASD-Introversion. This pattern hints that some older, more experienced teachers see themselves as more efficient, more self-assured but also more in-turned, which might help explain or confirm some of the mildly negative correlations between Age and COR-KU.

Preliminary interpretation: So far, the proposition has not been confirmed for the teacher self-report instruments. Teacher Age and Experience do appear to have some significant effects on the Kindly-understanding dimension of teaching behavior.

Proposition 4

Positive teacher characteristics will positively influence (predict) student behavior.

One preliminary analysis has been done that addresses this issue (Peck and Fox, 1977a). The ASD and VOT (and COR) teacher measures were regressed on the student measure of time-on-task behavior, based on individual classroom observations on a representative subsample within each class (the "special study" students). Both linear and quadratic terms of the teacher measures were included, to test for possible curvilinearity of relationship. This analysis then went on to regress student time-on-task behavior on regressed change in each of the outcome measures. This tested a three-link chain, presumably causal in nature. (See Proposition 8, below).

Of the nine teacher self-report measures tested for relationship to student time-on-task, in either linear or curvilinear fashion, three showed a significant effect (a meaningful number). Student time-on-task was positively related to teacher Efficiency, and to a moderate rather than extreme (reflected) Behavior score, for prudence, responsibility and courtesy, on the ASD. There was a curvilinear relationship of VOT: Attitude to Time-on-Task; the greatest facilitation was produced by teachers with the most highly positive scores.

Most of the analyses that test this hypothesis will be carried out in the next six months.

Preliminary interpretation:

So far, the evidence supports the proposition.

Proposition 5

Positive teacher characteristics will positively influence student outcomes.

Stage IV provides a preliminary test of this proposition, modified by any significant interaction findings in Stage V, and tested for possible class effects in Stage VII. Table 4 displays the teacher main effects, and the interaction of teacher variables with student pretest levels, which were discovered through the Stage IV analysis of the Austin Year One data.

The Derivation of Tables 4A and 4B

The following equation (1) represents the logical model utilized in Stage IV (Teacher Assessment) analysis. The elements of the equation correspond to the effects tested in this stage. Tables 4A and 4B display only those effects which met the following criteria, although all the effects were tested: (a) $p \leq .01$ and percent of variance explained $\geq 1.0\%$; and (b) $.05 \geq p > .01$.

$$Y = X + P^L + P^Q + X^L * P^L + X^L * P^Q + X^Q * P^L + X^Q * P^Q \quad (1)$$

where: Y - criterion variable - SET POST, SSC-A POST, SSC-POST, CTBS POST, SSI POST, P-H POST

X - Pretest of criterion variable

P^L - Linear component of Teacher Assessment predictor-VOT-Comp, VOT-ATT, ASD - ATRV, ASD-I-I, ASD-ATT, ASD-EFF, ASD-INT., ASD-ANX, ASD-BEH

P^Q - Quadratic component of teacher assessment predictor

$X^L * P^L$ - Criterion pretest-linear by Teacher assessment predictor-linear interaction

$X^L * P^Q$ - Criterion pretest-linear by Teacher assessment predictor-quadratic interaction

$X^{Q*}p^{L}$ and X^{Q*} - criterion pretest-quadratic by teacher assessment-linear interaction and criterion pretest-quadratic by teacher assessment-quadratic interaction

There are six outcome measures, with nine (9) teacher assessment predictors for each, making 54 main effect tests. Of the 54 tests, 19 were significant ($p < .05$). Although these significant effects are not necessarily independent due to the interrelationship among the outcome measures, as a whole, these effects are beyond chance.

To summarize the findings,one or more of the ASD scales predicted each of the outcome measures, except for SSC-A, at the .01 level of significance, and from one to three ASD scales additionally predicted every outcome criterion at the .05 level. A VOT scale predicted SSI change at the .01 level, and another predicted SSC-C change at the .05 level.

The findings further showed many interactions of student pretest with a teacher variable: one, affecting SSI; ten, affecting self esteem (P-H); four, affecting achievement (CTBS); five, affecting SET; and one, affecting SSC-C; (at either the .05 or .01 level)

To illustrate with one set of these effects, achievement (CTBS) gain was negatively affected by the ASD-Attractiveness score, positively, by ASD-Introversion; curvilinearly by a complex combination of pretest by ASD-Attitude factors, and by pretest by ASD-Behavior factors. It was also affected by ASD-Anxiety and by VOT-Coping, interacting with student pretest level of achievement (direction still to be determined).

A similarly sizable list of teacher effects, or teacher by student pretest effects, is shown in Table 4, for every criterion except SSC-A, which had only one (positive) effect, from ASD-Attitude.

Interpretation of the effects of each teacher characteristic will have to take into account any significant interactions that involve that variable, in the Stage V analysis (yet to be done on these measures). All of these effects must be further tested through Stages VI and VII to be sure they are validly identified.

The possible effects of teacher age, training, and experience on student outcomes also remain to be tested.

Preliminary interpretation: So far, this proposition is supported, for both teacher instruments; not strongly in the case of any one teacher variable, but pervasively across all of the teacher self-report variables.

Proposition 6

Positive teaching behavior will positively influence student behavior.

One preliminary analysis has been made, using class mean scores on the student measures, to test the relationship of the three teacher COR measures to student time-on-task behavior (Peck and Fox, 1977b). Systematic, organized teaching behavior proved to be quite strongly related to student time-on-task behavior in the Austin, Year 1 sample. (COR-KU and COR-SI did not show significant effects.) In addition, COR-KU, SO and SI correlated with the COR-Pupil behavior rating .54, .66 and .75 in Austin, Year 2. (Year 1 is to be analyzed shortly.) All other possible effects of COR and COS measures of teacher behavior on COR and COS measures of student behavior remain to be tested, in all four samples.

Preliminary interpretation: The proposition is confirmed, so far. (Presumably) positive teacher behavior does positively affect student behavior.

Proposition 7

Positive teaching behavior will positively influence student outcomes

Table 5 summarizes the number of significant teacher main effects (see Appendices B-4 and B-8), and the number of significant teacher-by-student interaction effects (TxP) (see Tables 6-11) on six student outcome measures (regressed change over the year), in the Austin Year 1 and Year 2 samples. While the number of significant main effects far exceeds the chance level, the large number of interaction effects (see Proposition 9, below) clearly requires that each teacher main effect be examined for possible modification by interaction effects that involve that variable.

For example, in the Year 1 sample, change in self-esteem (P-H) was affected by Teachers' Stimulating, inventive behavior; but the effect was actually opposite in different ethnic groups (see Table 2A; Appendix B-5, Table 10; and Appendix B-7, Table 5). Teacher COR-SI behavior was positively related to gain in self-esteem among Chicano students; it made little difference to the self-esteem of Anglos; but it actually had a negative effect on the self-esteem of Blacks. This finding did not reappear on the Year 2 sample, so it is clearly not true for all minority children; but it was substantial enough to look like a reality for the students who made up the Year 1 sample. In that group, Black children may have felt hopelessly overawed by highly stimulating, inventive teachers and thus lost self-esteem over the year.

In the case of the Student Evaluation of Teaching measure (SET individual score), such complications did not exist in the Year 1 sample. There, students' regressed change in the way they individually evaluated their teacher was simply and positively affected by teacher behavior: linearly in the case of

Table 5

Number of Effects on Student Outcomes of Three COR Teacher Measures
(Main, and in Interaction with student pretests)

Instrument	Year 1				Year 2				Same in Both	
	$p<.01, V\geq 1\%$		$p<.05, V<1\%$		$p<.01, V\geq 1\%$		$p<.05, V<1\%$		At least $p<.05$ both years	
	COR Main[1]	Inter.[2]	COR Main[1]	Inter.[2]	COR Main	Inter.[2]	COR Main	Inter.[2]	Main	Inter.
SET	3	0	0	11	To Do	1	To Do	7	To Do	0
SSC-A	2	3	1	8	"	0	"	3	"	0
SSC-C	2	1	1	4	"	1	"	3	"	0
Achievement	0	0	2	13	"	0	"	3	"	1
SSI	3	1	0	3	"	6	"	13	"	0
Self-esteem	0	1	0	14	"	6	"	5	"	4

[1]Class effects ruled out

[2]Class effects not yet ruled out

COR-KU, curvilinearly in the cases of COR-SO and COR-SI (see Appendix B-4). There were no interaction effects of these teacher variables with student variables that met the "strong" criterion ($p < .01$; % of variance = 1.0). Although there were eleven interaction effects that met the $p < .05$ "weak" criterion, none of them was replicated in Year 2, where all eight interaction effects were new and different from Year 1 (Tables 6A and 6B).

Preliminary interpretation: The general proposition is definitely confirmed in both Austin samples. The specific relationships, however, among a particular kind of teacher behavior, the several student entry characteristics, and a given student outcome, were greatly different in the two samples. Both the many, complex interaction effects and the different patterns in the two Austin samples at least provisionally imply one conclusion: It is neither valid nor educationally defensible to state any general "laws" by which an instructional prescription can be stated that will work equally well with different students, or equally well in the hands of all teachers. Unsurprisingly, too, it appears that the educational effects of such factors, and sets of factors, even when they are reasonably well-measured, may vary greatly because of other unmeasured circumstances which cause things to operate differently in different times, groups or places.

This is not a sentence of doom (unless it be on the long-time American proclivity for seizing upon this year's "new, quick, easy answer" to humanity's ills). Such an inference does suggest, however, that a great deal more research needs to be done to tell us why things operate differently on different occasions. Much future research will have to follow a multivariate, interaction model or we will arrive at findings which are not just overly simple, but are frequently incorrect in humanly important ways.

Proposition 8

Positive student behavior in the classroom will positively affect student outcomes.

One kind of student behavior, time-on-task, has thus far been tested for its effects on student outcomes in the Austin Year 1 sample (Peck and Fox, 1977b).

Student time-on-task showed six significant relationships, either linear or curvilinear, with regressed change on the outcome measures. Time-on-task had a positive effect on SSC-Attitude improvement. This showed up in a linear main effect but it also had a curvilinear component whereby the effect was even stronger for students who started out low or average in attitude.

Coping gains (SSC-Coping) were both linearly and curvilinearly affected by time-on-task, in a positive direction. The greatest gain in coping skill was associated with the highest level of time-on-task during the year.

Time-on-task had a positive influence on self-esteem (P-H). Working consistently during the school year led to maintained or improved self-esteem whereas low time-on-task behavior led to losses in self-esteem.

The standardized achievement test (CTBS) showed a curvilinear effect of time-on-task behavior. The highest time-on-task was conducive to the greatest gain. The largest difference in gain was between students who showed high and medium time-on-task behavior.

On three different outcomes, a high level of time-on-task made a much bigger positive difference in outcome than the medium or low levels of time-on-task appeared to produce. An average amount was better than a low amount of time-on-task behavior but the biggest and most positive effect was that of consistently high time-on-task activity.

The possible effects of all other measures of student classroom behavior await analysis.

Preliminary interpretation: The one piece of evidence to date confirms the proposition. It is a particularly important test, since evidence has recently been accumulating that time-on-task behavior has positively influenced learning gains, in other studies. The unique, added findings from this study are that it also has a positive effect on self-esteem (P-H), attitude toward life (SSC-A), and self-rated coping skill.

Proposition 9

Student outcomes will often be affected, not just by teacher or student characteristics taken alone, but by their interacting effects.

Logically, several kinds of effects can make it necessary to look at both student and teacher characteristics in order to foresee or understand a student outcome. In the first case, both a teacher and a student variable have independent effects; both must be taken into account in explaining the result. A simple example is the "best" main effects predictor equation given in Appendix B-8 for SET change: SET Post = SET PRE + SSI + COR KU + CLASS. Both student attitude toward school (SSI) and teacher kindliness (COR KU) made independent contributions to the students' final evaluation of their teachers. This is not, technically, an "interaction" effect. It does, however, show that the outcome cannot be explained adequately without taking both the teacher and student variables into account.

The next is a statistical interaction effect: T * Student Pretest (non-homogeneous slopes). In this case, the effect of the teacher characteristic is different for students who start the year at different levels on the criterion test. Thus, in Year 1, achievement gain was positively affected by teacher coping skill (VOT-C) but to different degrees depending on the student's entering achievement level (see Table 5).

Finally, a teacher characteristic may affect a student outcome differently, depending on students' entering level on a different student variable. This is the T * P effect most commonly referred to as an interaction. An example is given in Table 5 in Appendix B-5. In Year 1, attitude change (SSC-A) was positively related to teacher Systematic-organized teacher behavior (COR SO) at high or middle SES levels; but the effect of COR SO was actually slightly negative for low SES students. Thus, a statement based just on the

"teacher effect" means that Systematic-organized teaching improves students' attitude toward life, would be very untrue for low SES students in that Year 1 sample.

As indicated earlier, described in Tables 5-11, and further described in Appendices B-5 and 7, both kinds of statistical interaction effects appeared in the data from both years, in substantial numbers. Further, as Appendix B-8 describes, both teacher and student effects proved to be necessary elements in explaining a number of the outcomes.

The most interesting results are the substance of these effects, of course: precisely what teacher characteristics affect attitude change or achievement gain, in combination with what student characteristics. Before making any "final" statements of this kind, it is necessary to interweave the findings from all nine stages of the analytic system. There has not been time to do this yet, except for a trial effort with one or two variables.

Preliminary interpretation: What can be said now, with considerable confidence, is that teacher-student interaction effects have appeared on so many of the outcomes that the interaction model clearly is a necessity if one is to explain almost any one of these outcomes. It has taken very long, arduous effort, some of it necessarily of a trial-and-error sort, to invent a total analytic system that can identify and rigorously test such effects in a multiple-input, multiple-output research model. At the outset, no one could advise us on a way to get here because this had never before been attempted. As Mood observed, "My own struggles...have led me to believe that education is, at best, an order of magnitude more complex than economics (Mood, 1971)." Dunivant spent a year or more trying out the possible utility of econometric models for this study. In the end, we had to come to that same conclusion as Mood--

that the best econometric models only posed one or two simple questions, and answered them, in a greatly over-simplified way. A single, grand multiple regression equation may say a lot of certain kinds of things, but it is of no use in telling what are the differential effects of certain predictors, in specified combinations, for particular kinds of students.

At this point, the chief completed contributions of the study are the statistical system it has developed to meet this need, and the finding that there are too many significant interaction effects to be ignored. Their precise nature, and their practical implications, remain to be described.

Some teacher-pupil combinations that illustrate T * P interaction effects can be described, in a limited way, using the anecdotal reports drawn from the classroom observations and the interviews with the teachers. One such effect is illustrated by the following examples.

Proposition 10

Each outcome will be produced by a unique set of factors.

The generation of a total, final, "best" predictor equation for each outcome awaits completion of the Stage IX analysis. So far, the results of Stage VIII speak to this proposition insofar as concern the main effects, only, of some of the teacher and student predictors. Even at this simple level, using relatively few terms, the proposition is confirmed. To illustrate the diversity of the predictor models shown in Appendix B-8:

SET POST = SET PRE + SSI + COR KU + CLASS

SSC-C POST = SSC-C PRE + SEX + ETH + PH + BRS + SSI

SSS-I POST = SSI PRE + SEX + ETH + SET + SSC-A + P-H + BRS + COR KU + COR SI + CLASS

CTBS Post = CTBS PRE + SEX + ETH + SES + PH + BRS + CLASS

Some outcomes are affected by sex, some are not; some by SES or ethnicity, some not; some by teacher factors, some not; some demonstrated a class effect in the Year 1 sample, some did not.

Preliminary interpretation: So far, the proposition is supported by the evidence. It will almost certainly be further strengthened when all predictors, and their interaction terms, are added.

Proposition 11

Socioeconomic status will be positively related to students' levels on the entry measures.

In both years, the Stage V regression analysis found SES positively related to the pretest measures of achievement, peer-rated coping skill, and self-esteem. SES was not related to the self-report measures of attitude or coping skill (SET, SSI, SSC-A, SSC-C).

The most notable difference was in achievement. As Table 12 shows, there were large, systematic differences in both years. The specific scores of the population subgroups are shown in Table 13. In year 1, the upper SES group (approximately upper-middle and middle-middle class on Warner's scale) was about one half a standard deviation above the middle SES group (lower-middle and the upper half of the upper-working class) which, in turn, was half a standard deviation above the lowest group (the lower half of the upper-working class and the semi-skilled working class). In the following year, there was the same size gap between the upper and middle SES groups; an even larger (.75 S.D.) difference between the middle and lower SES groups.

As Table 13 also makes clear, there was a strong tendency for SES and ethnicity to be linked in the sample with complete achievement and SES data. Among Anglos, the frequency of their representation in the three SES levels was as follows. Year 1: High - 44%, Middle - 36%, Low - 20%; Year 2: High - 45%, Middle - 29%, Low - 26%. Blacks were distributed in Year 1: High - 12%, Middle - 30%, Low - 54%; in Year 2, Blacks' SES distribution was High - 13%, Middle 21%, Low - 65%. Chicanos in Year 1 were distributed: High - 12%, Middle - 34%, Low - 54%; in Year 2, High - 20%, Middle- 19%, Low - 61%. Nonetheless, contrary to our original expectations, by no means all of the lower achievement of the minority groups was accounted for by their socioeconomic disadvantages. (See Propositions 13 and 14, below, for a full discussion of this issue).

Close to five percent of the performance differences in Year 1 and nine percent in Year 2 (12%, with Missing SES data counted in), were attributable entirely to SES effects. Higher is better and lower is poorer; not just in monetary terms, but in skill mastery. The "missing data" term stands for students who supplied either no information or unusably vague or uncertain information about their families' occupations and years of schooling. This "missing" group had a much lower score on the achievement tests than the lowest of the SES or ethnic groups who did supply usable SES data. This was anticipated, from earlier research experience, and the Missing-data vector was accordingly tested in all phases of the analysis. In effect, it largely identified a group at the very lowest end of the SES scale, and of the achievement scale.

The SES performance differences on the achievement tests were mirrored by parallel SES differences in peer-rated coping skill (BRS-OR), and in self-evaluated self-esteem. The evidence on the latter point, drawn from the P-H and SSD instruments, was buttressed by relatively small but systematic SES differences, also, on the BRS-Self-Rating, in Year 2. (Year 1 data on this measure are not yet analyzed.) Thus, performance measures, classmates, and students themselves, all agree that there is a considerable linkage of academic skill, performance and adjustment, with family SES level. Learning does come hard to many children from the lower SES levels.

There is nothing whatever new in this finding. The International Study of Educational Achievement and the eight-nation study of Coping Styles and Achievement found this a near-universal pattern. Higher SES children perform a lot better in school in all countries (with the sole exception of Ljubljana, Yugoslavia, in the Cross-National Coping Study). If the public wants public education to equalize educational achievement outcomes for all children, a revolution will probably have to take place in social priorities, in funding, and in educational strategies.

One of the intents of this study is to cast some new light, if possible, on factors in children and in teaching that might point ways to move toward more nearly equal outcomes. So far, the evidence does confirm the expectation that the SES discrepancy in achievement is linked with a parallel disparity in self-esteem. That supplies one lead for further research into possible corrective actions. (See Propositions 7, 9, and 10). Further, as expected, these differences cannot be blamed entirely on "biased" or otherwise doubtful achievement measures. The differences in scores reflect differences in application, mental initiative, and willing acceptance of the human partners in schooling, as these are observed by classmates. Coping skill training is needed, too, if lower SES students are to increase their basic skills.

Table 12

SOCIOECONOMIC AND ETHNIC DIFFERENCES IN STUDENT PRETEST MEASURES (Austin)

($p \leq .01$, variance $\leq$ 1%)

YEAR ONE			YEAR TWO		
SET:		None	SET:		None
SSI:		None	SSI:		None
SSC-A:		None	SSC-A:		None
SSC-C:		None	SSC-C:		None
Self-Esteem: [ETH(Whole) p <.01, V 0.9%]			Self-Esteem: [SES(Whole) p .000, V 4.8%		
(P-H)	SES(Whole)	2.7%	(SSD)	SES(Partial) p .000, V 4.7%	
	SES(Partial)	2.3%			
Coping Skill (BRS-OR)	ETH(Whole)[1]	9.5%	Coping Skill	ETH(Whole)	4.0%
	ETH(Partial)[2]	6.4%		ETH(Partial)	1.6%
	SES(Whole)[3]	4.5%		SES(Whole)	4.3%
	SES(Partial)[4]	1.4%		SES(Partial)	2.3%
				SES(Missing)	2.1%
Achievement (CTBS)	ETH(Whole)	19.9%	Achievement (GMG)	ETH(Whole)	18.0%
	ETH(Partial)	11.2%		ETH(Partial)	9.0%
	SES(Whole)	13.4%		SES(Whole)	18.0%
	SES(Partial)	4.7%		SES(Partial)	9.0%
	SES(Missing)[5]	1.0%		SES(Missing)	2.6%
				ETH * SES(Missing)	2.0%

1. ETH(Whole) = Ethnic effect, not controlled for SES
2. ETH(Partial) = Ethnic effect, controlled for SES
3. SES(Whole) = SES effect, not controlled for Ethnicity
4. SES(Partial= SES effect, controlled for Ethnicity
5. SES(Missing) = Data on which to base SES score were missing, vague or unintelligible.

TABLE 13

SES and Ethnic Effects on Achievement (CTBS), Austin, Year 1

				MEAN		SIGMA		R(X,Y)
		No.	%	CT Pre	CT Post	CT Pre	CT Post	
ANGLO		753	71	518	514	75	87	.75
BLACK		149	14	416	425	73	69	.34
CHICANO		155	15	459	451	77	89	.72
SES	H	366	35	531	532	81	90	.75
SES	M	368	35	496	492	76	83	.72
SES	L	323	30	454	449	79	82	.64
ANGLO								
SES	H	329	44	539	541	76	86	.74
SES	M	271	36	510	503	73	81	.73
SES	L	153	20	489	476	65	79	.69
BLACK								
SES	H	18	12	420	447	80	80	.55
SES	M	45	30	430	427	69	61	.31
SES	L	86	58	408	419	72	69	.29
CHICANO								
SES	H	19	12	490	443	75	82	.72
SES	M	52	34	484	489	65	82	.71
SES	L	84	54	437	430	78	87	.70

SES and Ethnic Effects on Achievement (GMG), Austin, Year 2

				MEAN		SIGMA		R(X,Y)
		No.	%	GMG Pre-3	GMG Post-3	GMG Pre-3	GMG Post-3	
ANGLO		618	76	72	75	15	15	.86
BLACK		89	11	52	56	16	17	.86
CHICANO		106	13	61	64	16	18	.84
SES	H	314	39	76	80	13	13	.84
SES	M	220	27	69	72	14	15	.87
SES	L	279	34	59	62	17	17	.85
ANGLO								
SES	H	281	46	77	81	12	12	.82
SES	M	181	29	71	74	13	14	.85
SES	L	156	25	64	67	17	16	.85
BLACK								
SES	H	12	13	59	67	15	16	.91
SES	M	19	22	58	63	15	17	.90
SES	L	58	65	49	52	15	16	.81
CHICANO								
SES	H	21	20	73	77	10	12	.88
SES	M	20	19	63	67	14	18	.89
SES	L	65	61	56	58	15	16	.77

TABLE 14

SES and Ethnic Effects on Self Esteem, Austin, Year 1

				MEAN		SIGMA		R(X,Y)
		No.	%	P-H Pre	P-H Post	P-H Pre	P-H Post	
ANGLO		767	71	72.3	73.6	17	17	.70
BLACK		147	14	71.0	69.4	15	15	.64
CHICANO		165	15	68.3	69.1	17	17	.62
SES	H	379	35	75.0	76.3	15	15	.66
SES	M	374	35	71.4	72.1	17	17	.71
SES	L	326	30	67.5	68.0	17	17	.64
ANGLO								
SES	H	340	44	75.3	76.9	15	15	.66
SES	M	276	36	71.1	72.1	17	17	.74
SES	L	151	20	67.5	68.8	17	17	.63
BLACK								
SES	H	22	15	71.9	72.4	13	15	.72
SES	M	43	29	71.6	69.7	16	13	.43
SES	L	82	56	70.4	68.4	15	16	.72
CHICANO								
SES	H	17	10	71.2	68.7	19	16	.48
SES	M	55	33	73.1	74.1	14	15	.71
SES	L	93	57	64.8	66.2	17	18	.59

SES and Ethnic Effects on Self Esteem, Austin, Year 2

				MEAN		SIGMA		R(X,Y)
		No.	%	SSD Pre	SSD Post	SSD Pre	SSD Post	
ANGLO		646	77	23.6	24.1	6.6	6.7	.68
BLACK		87	10	22.5	23.1	6.5	6.6	.60
CHICANO		109	13	22.6	23.7	6.3	6.3	.54
SES	H	336	40	24.8	24.9	6.2	6.4	.65
SES	M	228	27	23.6	24.6	6.4	6.1	.63
SES	L	278	33	21.4	22.4	6.7	7.1	.65
ANGLO								
SES	H	301	47	24.6	24.8	6.3	6.3	.66
SES	M	189	29	23.6	24.5	6.4	6.4	.62
SES	L	156	24	21.6	22.4	6.9	7.4	.75
BLACK								
SES	H	15	17	25.6	24.6	3.8	7.4	.62
SES	M	19	22	24.7	25.4	5.9	4.5	.70
SES	L	53	61	20.7	21.9	6.8	6.7	.57
CHICANO								
SES	H	20	18	26.1	26.3	6.1	5.9	.57
SES	M	20	18	22.5	24.7	5.7	4.6	.72
SES	L	69	64	21.7	22.6	6.1	6.5	.47

Proposition 12

Student socioeconomic status (SES)will positively affect student outcomes.

In Year 1, SES did not affect Student Evaluation of Teacher over the year. Students were not biased in their judgment or attitude toward their teacher by their family SES level (i.e., the SET measure was free of this kind of bias). Neither were their changes in the other attitude measures (SSI, SSC-A) affected by their SES background.

Their achievement gains, however, were different at the different SES levels; but this effect was mixed interactively with an ethnic effect, and was therefore affected by the large difference in pretest means of the ethnic groups, and also by the positive correlation of the pretest with SES (See Appendix B-1, and Appendix B-2, Table 2). Under these circumstances, even the special analysis that was carried out has uncertain validity. Conditions are not right for covariance analysis when pretest differences are so large. In the near future, a separate analysis must be run on each of the SES levels, and each of the ethnic groups. Meanwhile, the data seem to indicate that achievement gains went up as SES went up, among Anglos; there was only the slightest tendency in this direction among Blacks; and, among Chicanos, the effect of SES was actually a little negative -- the higher status Chicanos gained less than those of lower SES. Pooling all of the students gave a positive effect of SES on achievement, but only because the large effect among the Anglos outweighed the smaller negative effect among the Chicanos.

All of these effects were further complicated by the effect of the pretest level on gain; another indication that the new analyses are needed.

Higher SES, in Year 1, did have a positive effect on change in self-reported coping skill (SSC-C), but most of this effect was actually due to

ethnicity, which was significantly associated with SES. It was not so much high SES that made for this greater gain, but Anglo ethnicity, especially as compared with the Black students. (See Proposition 14, below.)

SES had an effect on change in self-esteem (PH) in Year 1: higher status meant more positive change.

In the Austin Year 2 sample, however, SES had no significant effect on any of the outcomes except achievement. Higher SES was associated with greater gain. The "rich got richer (more knowledgeable) and the poor got poorer," in this one respect, in both years. This analysis, too, must be carefully examined and almost certainly rerun separately for the different ethnic and SES groups. Otherwise, the Year 2 sample behaved very differently from the first year sample, in this analysis as on many of the other analyses. The same antecedent variables (SES, in this instance) did not have the same effects on the outcomes, in the two samples.

Preliminary Interpretation: Students generally were not biased toward school or teachers by their SES, although it appeared to make a big difference in their achievement. Further analysis is necessary to clarify this, because of the large SES and ethnic differences in achievement pretest scores. One of the principal findings of this aspect of the study, not surprisingly, is the interplay of SES and ethnicity. It is particularly important with these variables, to partial out the variance attributable to each, or their combination. It is vital to know whether an effect is due to family SES or to ethnic culture.

Why the SES effects were not the same in the two samples poses a challenging question for further inquiry into the social dynamics of these groups.

Proposition 13

There will be significant ethnic differences on the student entry measures.

There were two instruments that showed substantial ethnic differences on the pretest, in both years: the standardized achievement tests (Table 13), as all previous evidence would predict and, less expectedly, the peer rating of academic coping skill (BRS-O). Indeed, in Year 2 (Year 1 data is not yet analyzed on this point), even the self-rating on the BRS (BRS-S) showed the same ethnic differences as did the peer ratings. On all three measures, Anglos scored highest, Chicanos intermediate, and Blacks lowest. This congruence, coupled with the fact that the peer ratings proved to be free of ethnic bias (Peck, Clements & Weber, 1977) suggests that there are genuine differences in level of performance and academic coping skill among the three ethnic groups. The difference in standardized test scores appears to reflect this to a substantial degree, not just presenting a distorted picture because of content-bias or other culturally unfair characteristics. The Black and Chicano students who were mixed in with the Anglos all agreed that the intellectual skill, application and cooperativeness of the three ethnic groups differed in the same order as their test scores; and so did their self-evaluations on these characteristics. It looks as though the coping skills represented on the BRS may identify those places where minority children need help, and where they <u>can</u> be helped (Spaulding, 1970), in ways that could improve their mastery of the knowledge and skills tapped by the achievement tests.

Three other instruments, assessing attitude toward the specific teacher (SET), attitude toward school in general (SSI), and general attitude toward life (SSC-A), were not affected by ethnicity. They were free of ethnic

bias on the students' part, it could be said. This is extremely important, for several reasons. Most important, it indicates that the minority children did not enter these sixth grade centers with any more negative expectations or disaffected feelings about school than the majority children. In this respect, at least, their experience seems to have led them to feel equally accepted in school and equally acceptant toward it. They also described their own coping skills at the same level as the majority children (SSC-C).

In the Year 1 sample, ethnic differences were present in initial self-esteem (PH; Table 14) but these disappeared when SES differences among the ethnic groups were partialled out. It was low SES, not minority membership, that made for lower self-esteem; although most of the minority students were of lower SES and therefore lower self-esteem. In Year 2, no ethnic difference appeared in self-esteem (SSD).

Preliminary Interpretation: There seem to be real differences in coping skills that at least partially explain the ethnic group differences in academic achievement. Even by sixth grade, however, the minority children were no different from majority children in their attitudes toward school and teachers. The Year 1 finding of a difference in self-esteem vividly illustrates the need to partial out SES: it was low SES, not ethnicity, that made the difference.

Proposition 14

Ethnicity will affect student outcomes.

Ethnic membership had no biasing effect on students' changes in attitude toward school (SSI) or teacher (SET) over the year. They neither lost the unprejudiced approach with which they began the year, nor were more vulnerable to disaffection or negative attitudes if they were Black or Chicano. Ethnicity did affect change in general outlook (SSC-A) to a modest degree in the Year 1 sample; minority children were apt to grow a little less optimistic and positive. In the Year 2 sample, no such ethnic effect appeared on SSC-A. There was also an ethnic effect on change in self-esteem (PH) in Year 1: minority children were more apt to lose a little ground; but SES accounted for most of this (see Appendix B-2). No ethnic effect on self-esteem (SSD) change appeared in Year 2.

In both years, there were ethnic differences in achievement gain (Table 13). This is subject to the same uncertainty as was discussed under Proposition 12. Large differences in the pretest scores of the ethnic groups require that further analysis be done of changes in the separate groups.

In both years, too, there was an ethnic effect on change in self-described coping skill (SSC-C). Anglos fared best; Chicanos, next; and Blacks showed the least favorable change.

Preliminary Interpretation: Minority membership did not appear to cause less favorable changes on most of the attitude measures; nor did they have any appreciable problem with self-esteem when SES effects were taken out.

There was a disadvantage to minority students in their picture of their own coping skills: it went down, over the year, in both samples. There also seemed to be a widening of the achievement gap, over the year, especially between Anglos and Blacks; but this must be tested further.

Proposition 15

Many of the specific relationships described under Propositions 1-14 will prove to be curvilinear.

As expected, curvilinearity has turned up to a significant degree from Stage IV through Stage VII, wherever quadratic terms were included (See Appendix B4-7). Some examples are presented graphically in Figures A, B and C in Appendix B-5. Figure B, for instance, which corresponds to Table 15a in that appendix, shows that low Teacher Kindliness (COR KU) did not have much effect on school-attitude change (SSI) among students who started the year with low self-esteem (PH); and even teachers medium in COR-KU did little good for such students, whereas teachers high in COR KU had a measurably better effect on their attitude. For students originally average in self-esteem, the degree of Teacher Kindliness made the most difference (the more, the better). Highly self-esteeming students were not all that much affected in their attitude, but insofar as they were, it was chiefly a more negative influence by low COR KU teachers.

Preliminary Interpretation: Thus far, the proposition is confirmed. More is not always better. There are "threshold effects" and "ceiling effects" at work in many relationships. Life does, indeed, throw us quite a few curves -- metaphorically, statistically, and in many very real, perceptible ways. It is best to allow for this in examining all kinds of data.

Proposition 16

Each student instrument will show a significant degree of validity, in up to seven ways.

The seven kinds of validation built into the design are described above in the Analysis Section. The detailed results for each instrument (so far as they could be assembled in time for this progress report) are given in the manuals in Appendix A. One way to summarize this sizable array of relationships is to note that the significant proportion of positive findings, to date, on Propositions 1-14, also add up to substantial validation of several kinds, for all of the student measures. The findings on Proposition 1 show cross-instrument validity among the pretests. The data for Proposition 2 show substantial predictive validity, when change from pretest is the criterion. The evidence available on Propositions 4, 5, 7 and 8 show the kind of consequent validity that was predicted. Propositions 11-14, which predicted SES and ethnic effects, have so far shown only partial confirmation of the general proposition that all student measures would be affected; but this is not a test of their validity as measures of the qualities they are supposed to measure. Propositions 9 and 10, on the other hand, do directly assess whether particular student characteristics have the effects they logically should show if the instruments that purport to measure them are valid.

The only really satisfactory way to look at the results, of course, is to take one instrument-variable at a time and see if it is validated by those particular antecedent, contemporary, and consequent variables to which it should show significant relationships if it is validly measured (and if <u>they</u> are validly measured).

This kind of research is just the kind that is projected in the proposed Longitudinal Study of Coping (See Future Research, below.)

<u>Preliminary Interpretation</u>: On the whole, the evidence to date confirms the proposition. The measures in the network of instruments are proving significantly

related, in predicted directions. The size and strength of the validity evidence varies from one instrument to another. Ultimately, the validity of each measure will be separately summed up.

Proposition 17

Teaching behavior will be validly measurable.

So far, the kinds of validation tests that have been carried out have yielded positive results. The estimated reliability of 53 Austin teachers' scores on Ryans' three factors, computed from the independent ratings of four observers on two or three occasions, was .97, .98 and .98 for COR KU, COR SO and COR SI, respectively. This evidence demonstrates a very high degree of reliability of measurement across both raters and occasions. When 28 of these teachers were rated again in the second year by a partially new team of observers, in different subject-matter classes, with new students, the correlations of COR with the first year ratings were: KU .77, SO .74, and SI .58. Taken just as a validity test, this implies that the COR ratings from which the factor scores are derived are quite valid measures of these aspects of teaching behavior.

The factor structure of the first-year ratings almost exactly reproduced that found in Ryans' original study (Ryans, 1960), although the observers did not know what factors had been extracted from the scales they were taught to rate. This is evidence of the validity of the construct system represented by Ryans' three factors.

The student rating of teachers (SET class mean score) supplied an independent measure. In Year 1, this correlated with COR-KU .68, COR-SO .54, and COR-SI .47. In Year 2, SET scores generated by new classes of students correlated with COR-KU .54, COR-SO .45 and COR-SI .38. Although the student bodies were entirely different in the two years, when the teachers' first year SET scores were correlated with their second year SET scores, the result was .68, a quite impressive validity coefficient for SET.

Insofar as pupil behavior, as measured by the COR-Pupil rating, might be considered a measure of teacher impact, and if similarity of impact on different classes might be taken to imply the validity of the assumption, the correlation of COR-Pupil of .68, across the different classes in the two years, further suggests that COR-Pupil tends to be a valid index to something in teaching behavior that has a predictable effect on students. To rule out the possibility that this rating of pupil behavior might be due primarily to the similarity of the pupils in the two classes per teacher, one rather crucial test of similarity was made. The classes' mean scores on the achievement pretests and posttests were compared across the two years. The correlations were essentially zero. The two classes each teacher had in succession did not resemble one another in achievement. Finally, the classes' pretest achievement scores were uncorrelated with their teachers' COR scores in either year.

Preliminary Interpretation: The evidence examined thus far indicates that Ryans' three COR factors, Kindly - understanding; Systematic - organized and Stimulating - inventive are valid measures of the qualities they purport to measure. The COR-Pupil score appears to be a valid estimate of teacher impact on students, whatever qualities of students may also enter into it. The Student Evaluation of Teaching (SET) shows the same, quite high degree of validity. Other measures remain to be assessed. So far, the proposition is confirmed.

Proposition 18

Individual teaching behavior will prove relatively stable across time and across classes, when the measures possess a suitable level of generality.

The reasons for selecting Ryans' observation system were described earlier. The one most relevant here is that his medium-inference scales "smooth out" differences in the details of teaching behavior which are the data in low-inference coding systems. Phenotypic variations due to subject-matter differences, differences in administrative policies, or other situation-specific constraints, have entirely obliterated any appearance of consistency in teacher behavior in a good many studies that used only low-inference measures. Some previous research on teaching and research on other occupations (Roe, 1953, 1956; Stern, Stein & Bloom, 1956; Super, 1957, 1970; Peck & Parsons, 1956; Peck, 1959, 1960; Siegelman & Peck, 1960; Ryans, 1960; Holland, 1973; and Ryans, 1972) has demonstrated that occupational behavior is a direct, lawful extension of relevant personality characteristics of the individual carrying out the job. If this is true of teachers, it should show up in the form of a significant degree of consistency in individual teaching behavior across different occasions, even across different classes of students, different subject matter, and in successive years.

Certain parts of the evidence presented for Proposition 17 are even more centrally applicable to this proposition. Indeed, there is an inevitable relationship between the two propositions. If there were no generalized consistency in individual teaching behavior, but only content-specific patterns, there could be no appreciable correlation of teaching measures across occasions or classes. Validity could still be found for a context-limited pattern by comparing the assessments made by observers and pupils, e.g.; but there would be little correlation across different occasions and subjects

Each of 53 Austin teacher's average degree of CORKU, CORSO, and CORSI classroom behavior was taken by averaging the ratings of four observers on either two or three occasions. The estimated reliability of this average score was .97 to .98 on the three COR factors. This could not happen unless there were high interjudge reliability on any one occasion and also a high degree of consistency in the behavior of individual teachers across occasions, as measured by the Ryans factors. However, the chief evidence for stability comes from the study of the 28 Austin teachers for whom complete data were available in both years. These teachers must have been behaving similarly in their different subject-matter classes, in two successive years, or there could not have been the quite high correspondence in SET ratings by their two different sets of students ($r = .68$); or the correspondence of Year 1 observer ratings with the ratings made by a partially different team of observers in Year 2 (COR-KU, .77; COR-SO, .74; COR-SI, .58). More inferentially, consistency was also indicated by the similarity of their pupils' behavior in the two years (COR-P, .68) if COR-Pupil be viewed as a measure of teacher impact, as discussed under Proposition 17.

Preliminary Interpretation: The proposition is quite impressively confirmed by the analyses done up to now on the two Austin samples, using the COR and SET measures. The Year 2 COS codes, Interviewer Perception ratings, and How the Class Makes Me Feel data remain to be analyzed, as do all of the Kentucky data. So far, the findings indicate that most teachers do tend to be quite consistent in the dimensions of behavior represented by the COR factors and the SET factor. We know, from all of the observational and interview evidence, of two individual instances where the teacher's classroom behavior changed substantially from one year to the next, while being quite consistent within each year. If these two cases were taken out of the comparison, the consistency of behavior of the other teachers would unquestionably show itself even higher than the correlations on the total group seem to say.

Proposition 19

Correction for instrument reliability will improve the accuracy of the analyses and hence the validity of the findings, particularly as they are interpreted in terms of the properties at issue, not just their psychometric representations.

For reasons already explained, no such correction has proved valid, itself, up to this time. We will be exploring the applicability of Joreskog's new method, this next year, and will try it out if it does look valid, applicable and economically feasible to use with this kind of multivariate data set.

Preliminary Interpretation: At present, the proposition remains Not Proven.

Proposition 20

Findings will replicate across samples.

As has been described under the earlier propositions, certain kinds of findings have turned up the same in the Austin Year 1 and 2 samples. Proposition 1, about the positive intercorrelation of student pretests, was replicated. So was Proposition 2, in its general form, about the significant effect of a pretest on change in the other outcome dimensions; though differences in specific relationships appeared. The failure of the teacher self-report instruments to predict COR scores for teaching behavior was the same in both years (Proposition 3). Propositions 11-14, about SET and ethnic effects, turned up very similar findings in both years. The chief exception was the non-appearance of SES or ethnic effects on change in self-esteem, in Year 2, where there had been some effect in Year 1.

Propositions 4-8, and 10 have not yet been tested with the Year 2 data.

Proposition 9 was definitely confirmed in its general sense: teacher-pupil interaction effects turned up in considerable numbers, in both years. However, their particular nature was almost entirely different in the two different sets of classes. Either this represents chance fluctuations so strong as to somehow override the statistical tests of significance; or more probably, it points to important though unmeasured differences in the two samples which did, in fact, alter the way that specific teacher and pupil factors affected pupil outcomes in these two different sets of students. There was some difference in the teacher set, too, which will be removed in a subsequent analysis that includes only those teachers present in both years.

However, in real life, the student body changes each year, at any given grade level in any school. In most schools, there will be some changes

in the teaching staff, too, in most years. Consequently, the analysis that has been done up to this point best represents typical school populations. The non-replication of many detailed interaction effects in the two years may therefore have extremely important implication for educational practice, and for most-needed future research.

Proposition 15, about curvilinear relationships, was generally confirmed in both samples. Again, important differences appeared which have yet to be examined, in the specific places where curvilinearity appeared in the two years.

Propositions 16 and 17, about instrument validity, seems to be confirmed similarly, so far as the analysis of Year 2 data has proceeded.

Proposition 18, about the consistency of individual teaching behavior, seems definitely confirmed in both Austin samples, and across them.

Proposition 19 has not yet been properly testable.

Preliminary Interpretation: It is proving possible to compare findings across samples, to test not only the general propositions but the many specific variable-relationships subsumed under each one.

Where generalizability appears across the two Austin samples, and if it also occurs when the Kentucky samples are included, some of the findings may suggest "universal" uniformities. These will have to be tested with other samples that differ in age level, culture, region, community characteristics, school policies, curriculum and other important respects.

Not surprisingly, the substantive findings that have been replicated so far largely seem rooted in the relative consistency of individual behavior through time, whether of teachers or students.

So far, though, it is equally apparent that generalized instructional prescriptions that are based on presumed "main effects" of either teacher or

pupil characteristics, operating independently of one another, probably seldom can be justified. There is too much curvilinearity and there are too many teacher-student interaction effects for such "universal" prescriptions to be valid, without a great deal of tailoring to particular teachers and particular students.

What kind of tailoring? The great variation in specific antecedent — (presumably) consequent relationships in the two Austin samples raises the strong probability that many more factors need to be taken into account than is done even in this relatively large, multivariate design. If this is true, it may further mean that it will prove to be a practical impossibility to find out enough about a new student body, and its teaching staff, in time to pre-select instructional strategies which the outcomes nine months later would show to have been best in that particular year, place, and human group. If this were to be demonstrated, it would appear to lead to several conclusions:

(1) A great deal of additional multivariate research will need to be done, over a considerable time, with many samples, if we are ever to arrive at a sound array of alternative instructional prescriptions and a valid set of decision-making rules for telling when to invoke which prescription. This is somewhat akin to saying that education will remain a guessing-game until it accumulates roughly the same amount and quality of empirical research that medical practice has received over the last hundred years.

(2) Future research will have to include measures of many kinds of relevant variables that it was not possible to include in the present design. The logical and analytical system developed for this study could accommodate such additional variables. The difficulty of cross-linking many more and different classes of variables -- intrapersonal, contextual, cultural, or whatever, is scarcely to be minimized.

(3) Not only before such research is done, but probably forever, the individual teacher will have to retain the task of making accurate diagnoses of children's individual circumstances and learning needs, the task of wisely and knowledgeably prescribing the "most favorable available" teaching-learning procedures, and the task of following through alertly and responsibly. This may be the only way ever to get the job done in real-time. No partially-computerized system of data collection and processing is likely to have either the logical power, the speed, or the affordability to do this, in the foreseeable future. Computers are likely to become an increasingly useful component in some facets of education -- if the public will pay for them; but they can't do the instructional decision-making. The art of teaching, like the art of medicine, will have to remain just that: an art, calling for the greatest possible knowledge, wisdom, compassionate concern and dedicated responsibility.

(4) If that be true, it becomes even more important than ever to invest research talent, money and community-wide, profession-wide cooperative effort in improving the accuracy, fairness and completeness of our procedures for evaluating teachers, for training them in the individually-most-appropriate ways, and for giving them self-assessing, self-correcting ways of continuing to improve their effectiveness. (See Future Research Extensions below: Teacher Follow-Up Study; Teacher Evaluation Research.)

All of this is speculative; but not entirely so. Such ideas can only be put forward as "conclusions" when all the data, on all samples, are completely analyzed. Nevertheless, the single non-replication of so many teacher-student effects in just the two Austin samples is pretty strong evidence that something like this chain of reasoning must be investigated. For now, though,

this series of hypothetical inferences is set down solely in order to invite critical reactions to these potential interpretations and to invite alternative interpretations of the data.

EXTENSIONS INTO FUTURE RESEARCH

Direct Continuations of Present Leads

Several alternative lines of further inquiry, none of them mutually exclusive or unrelated, suggest themselves. It would be valuable to replicate the design of the present study using new, different samples from other parts of the U.S., other ethnic groups, other age levels, etc. Such research will probably be contingent, however, on resolving the present, very serious deterrent to full, representative sampling of students that arises out of federal regulations requiring an extreme form of "informed parental consent" and other well-intended protections against human risk.

It seems essential, moreover, to conduct intensive, longitudinal case studies of students selected to represent some of the major educational issues identified in the present study. This would look for causes of their behavior, in ways that could lead to much more tightly specified instructional prescriptions, tailored to the thoroughly understood needs and reaction patterns of each type of student. Later on, experiments in classroom instruction might be set up, to test instructional strategies that look most promising, judging by the results of current and next-step research. This sounds easier and simpler than it usually turns out to be in the real world of schools. The schools are so swayed with both reasonable and unreasonable pressures to be all things to all peoples' children, never failing any student, yet never failing to educate every student to a high degree of skill, that any experiment is a real intrusion. Nonetheless, teachers are typically so eager for anything that may do a better job for their students that they will usually cooperate when there is even a fighting chance that experimentation may pay off. Indeed, the problem is often to keep an experiment clean by not having additional "enrichment" added, part way through, which might indeed be good for some students. The

short-run, wholly proper concern for student welfare, can genuinely conflict with concern for testing the effects of a specific instructional practice that might yield better results for all future students, in the long run.

Another necessary step is to test, experimentally, alternative procedures and materials for training pre-service and in-service teachers, drawing on the insights, methods, and usable findings from this study and other applicable research.

Such further work could range from small scale projects to large research programs. Three specific research programs have been designed as next-step extensions of the current work. It would be highly desirable that the first two proceed in parallel, so that the findings from the Longitudinal Coping Study could be incorporated as soon as possible into the instrumentation and pre-service training of the Teacher Follow-Up program (as well as into the projected Discipline and Dropout study by Sikes and Powell; see below). The three studies are:

1) a longitudinal study of coping skill development and adult career performance;
2) a longitudinal teacher follow-up research, in collaboration with Sandefur and others at Western Kentucky University; and
3) teacher evaluation research. A more detailed description of each proposed study follows.